

**A methodological investigation of non sampling
error: interviewer variability and non response**

Richard D. Wiggins

**A thesis submitted for a degree of
Doctor of Philosophy**

The London School of Economics

July 1990

UMI Number: U058633

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

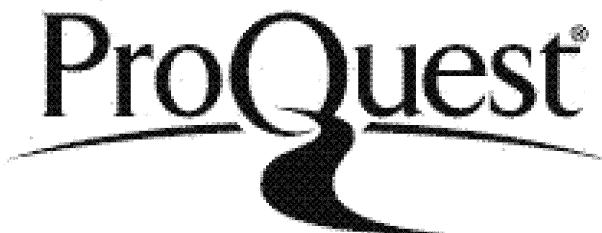


UMI U058633

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code.



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

THESIS

F
6807

x21116200x

Abstract

Two principal sources of error in data collected from structured interviews with respondents are the methods of observation itself, and the impact of failure to obtain responses from selected individuals. Methodological strategies are developed to investigate practical ways of handling these errors for data appraisal. In part one, the differential impact of each of a group of interviewers on the responses obtained in two separate epidemiological studies is examined. Interviewer effect is measured and its impact on the interpretation of individual responses, scale scores and modelling is shown. The analysis demonstrates that it is possible to achieve four objectives with slight modification of survey design. First, estimates of precision for the survey results can be improved by including the component due to interviewer variability. Secondly, items with high sensitivity to interviewer effect can be identified. Thirdly, the pattern of distortion for different types of items can be discovered. Replicate analyses appear to indicate that deviations between interviewers are not always consistent over time. Fourthly, by means of 'variance component modelling' the presence of interviewers on the interpretation of linear models can be evaluated. These models are used to show how interviewer characteristics may be used to account for variation in the responses.

Part two establishes an evaluative framework for the systematic review of interviewer call back strategies in terms of nonresponse bias and the costs of data collection. Use of an 'efficiency index', based on a product of 'mean square error' and cost for items in a survey of occupational mobility provides a retrospective evaluation. The empirical evidence had important practical consequences for fieldwork. The possibility of alternative call-back norms and the relative efficacy of appointment versus non-appointment calls is shown. The methodology develops from a review of adjustment procedures for nonresponse bias and models for survey costing. Logically, the methodologies for the three empirical investigations could be combined into an appraisal for a single survey. Only lack of resources inhibited such an outcome.

KEYWORDS: survey methodology; nonsampling error; interviewer effects; nonresponse bias; survey costing; callback strategies; variance component modelling, VARCL.

Acknowledgements

This thesis is largely empirical. Three major surveys support its illustrations. The work on the occupational mobility survey was initiated at the Department of Sociology at Queen's University, Belfast, where the author was a sampling consultant during 1973. I would like to express my gratitude to Professor J Jackson and Bob Miller of Queen's. The work on aircraft noise and physical handicap was initiated at the Institute of Psychiatry and the Department of Community Medicine at St Thomas' Hospital respectively whilst the author was a medical statistician during 1975 to 1979. I would like to acknowledge the support and encouragement received from Professors Michael Shepherd, Walter Holland and Dr's Alex Tarnopolsky and Donald Patrick. In addition, thanks to Yoga Sittampalam.

All of the interviewing was conducted by Social Community Planning Research. Special thanks to Jean Morton-Williams, Richard Stowell, Jane Cook, Jane Ritchie and Denise Lievesley, all of SCPR, and, of course the interviewers who must remain anonymous.

Computing help was essential. I am indebted to Margaret Menari, Frances Blomeley, Maria Tuck and Duncan Roweth at the Polytechnic of Central London. Special thanks to Dr's Gavin Ross with M.L.P and Nick Longford with VARCL.

Throughout the fifteen years of varying personal commitment my supervisor, Colm O'Muircheartaigh has been a valuable source of inspiration and encouragement. I hope I prove worthy of his continued belief in my abilities. Thank you, Colm.

And, of course, behind the scenes there has been encouragement from family and friends. In particular, Jack for suggesting the merits of planning, Kate for her optimism and Jenny for never doubting I'd get there! Thanks too to Mel Slater for being a thoughtful friend, Charlotte Wynn-Parry for providing perspective and Yvonne Goldstein for her remarkable skills with the Apple Macintosh.

..... it was one of those occasions when Chance having blindly prepared for a dramatic climax through aeons of time, seems to emerge like a palpable presence, out of the criss-cross accidents of its whirling possibilities and to survey its achievement

John Cowper-Powys
Weymouth Sands

A methodological investigation of non sampling error: Interviewer variability and non response

Contents:		Page
Chapter 1	Introduction	19
Part One	Interviewer Variability	28
Chapter 2	The assessment of interviewer effect	30
Chapter 3	A structure to investigate interviewer variability: the linear model	62
Chapter 4	Contextual issues: multivariate assessment and repeat measurement	118
Chapter 5	Specific survey conditions and the context for evaluation : the Aircraft Noise Survey and the Physically Handicapped Survey	127
Chapter 6	Resume of the evaluation strategy for interviewer variability	145
Chapter 7	An empirical evaluation of two interviewer variability experiments	154
Chapter 8	Variance component modelling of interviewer effects	204
	Conclusion for part one	225

**Part Two The impact of non response bias
and a global evaluation of the extent
of its' effect**

	Page
Introduction to part two	238
Chapter 9 A review of the effect of non response and strategies for exploring its' impact on survey estimates	241
Chapter 10 The interplay of non response bias, costs and other sources of error	289
Chapter 11 Specific survey conditions for the empirical evaluation : the occupational mobility survey	327
Chapter 12 A resume of the strategy to evaluate the impact of non response in a survey with call backs	332
Chapter 13 A global evaluation of non response bias in terms of mean square error and survey costs	339
Conclusion to part two	361
Chapter 14 Summary and speculation	367
Bibliography	384
Appendix	411

List of figures and titles

Chapter 1:

Figure

- 1.1 Total survey error
(Kish 13.2.II, p. 521)

Chapter 2:

Table

- 2.1 Hypothetical willingness to report medical conditions in relation to percentage of cases actually reported; Cannell p. 547
- 2.2 Illustration of specific interviewer behaviours; Cannell p. 581/2
- 2.3 Zone allocations (taken from Deming, p. 250)
- 2.4 Classification of major studies of interviewer variability by design schema proposed by Hansen, Hurwitz and Madow (1953)
- 2.5 Some empirical results for values of roh (taken from Kish, 1962)
- 2.6 Multiplier effect of interviewer variability on the variance of the sample mean
- 2.7 Some more results for roh from recent surveys

Figure

- 2.1 A model of the interview; Kahn and Cannell
- 2.2 A model of the interview situation; N M Bradburn
- 2.3 Factors affecting the interview; Cannell p. 539
- 2.4 A model of information processing in the interview; NCHS Report p. 53

Chapter 3:

Table

- 3.1 The curves of binomial variation (taken from Kish)
- 3.2 Estimators of μ , and F-statistics for testing $H: \mu = 0$, in three different restricted models
- 3.3 Estimable functions in the 2-way nested classification
- 3.4 Equivalent expressions for sums of squares in the analysis of variance of the 2-way classification
- 3.5 Suggested conclusions according to significance (sig) and non-significance (NS) of F-statistics in fitting a model with two main effects (α 's and β 's (tables 3.2 to 3.5 all from Searle, 1971)

Figure

- 3.1 A guide to the way theory can help determine strategies to investigate interviewer effect
- 3.2 Refusals and nonrefusals for 9 interviewers (taken from Deming, 1960)
- 3.3 An illustration of how regression slopes may vary in variance components analysis
- 3.4 A pathway for the exploration of interviewer effects

Chapter 4:

None

Chapter 5:

Table

- 5.1 Response rates by interviewer
- 5.2 Workloads analysed in variance component modelling
- 5.3 The experimental interviewers
- 5.4 Completed interviews by disability status
- 5.5 Workload character: Univariate analysis of selected socio-demographic variables
- 5.6 Functional limitations categories for disability survey
- 5.7 Subsamples analysed for combining and separate analyses in the physically handicapped survey
- 5.8 Interviewer characteristics

Figure

- 5.1 The West London 1977 survey of psychiatric morbidity
- 5.2 The Lambeth Health survey: location of interviewer effects study

Chapter 6:

Table

- 6.1 Statistical evaluation of roh (taken from Kish, 1962)

Figure

- 6.1: A three level hierarchy for time and interviewer
- 6.2: Resume of hierarchies used in variance component modelling in Chapter 8

Chapter 7:

Table

- 7.1 Distributions of values of \hat{p} for 41 questionnaire items.
- 7.2 Values of \hat{p} , appropriate standard error and variance multiplier for the GHQ and Annoyance scale scores
- 7.3 An illustration of variance estimation for the GHQ score
- 7.4 Degree of expressed annoyance with aircraft noise by interviewer
- 7.5 Log linear models for two way table of annoyance by interviewer
- 7.6: Univariate analysis of F.L.P. items
- 7.7 Intercorrelations of values \hat{p} for F.L.P. items between waves 1 and 2
- 7.8 Univariate assessment for items comprising the sleep and rest scale
- 7.9 Univariate assessment of items comprising the recreation scale
- 7.10 Interviewer effect for F.L.P. category means
- 7.11 Reliability after discarding successive items (from McKennell, 1974)
- 7.12 Using individual \hat{p} values as a criterion for eliminating items from a category scale: Sleep and Rest

7.13 Using individual $\hat{\rho}$ values as a criterion for eliminating items from a category scale:
Recreation

7.14 Multivariate analysis of variance: waves 1 and 2
(Wilks Lambda)

7.15 Interviewer effects matrices for sleep and rest items: (a) wave one, (b) wave two

7.16 Principal component analyses on matrices of interviewer effect for sleep and rest category:
(a) wave one, (b) wave two

7.17 Principal component scores for the first two components identified in table 7.15:
(a) wave one, (b) wave two

7.18 Clusters of items identified by PCA of "effects" matrices on sleep and rest items:
(a) wave one, (b) wave two

7.19 Correlations of "effects" between waves for sleep and rest items:
(a) Interviewers across items between waves
(b) Items across interviewers between waves

7.20 A comparison of "baseline" means used to estimate interviewer effects

7.21 Principal component scores for wave 1 "effects" on sleep and rest: unrotated and rotated solutions

7.22 Correlations of principal component scores with effect ($\sum n_i \alpha_i = 0$) for wave 1: sleep and rest

7.23 Selected interviewer effects for three sleep and rest items

7.24 Loadings matrix for wave one interviewer effects based on eleven interviewers in both waves

7.25 Loadings pattern specified in alternative confirmatory analysis for wave two interviewer effects correlation

Figure

7.1 Relative cumulative frequency distribution for \hat{p} on annoyance scale items

7.2 Relative cumulative frequency distribution for \hat{p} on 29 GHQ scale items

7.3 Relative cumulative frequency distributions for p on 135 FLP scale items (wave 1)

7.4 Relative cumulative frequency distributions for p on 135 FLP scale items (wave 2)

7.5 Plot of principal component scores based on unrotated solution in table 7.20

7.6 Plot of principal component scores based on rotated solution in table 7.20 (plus transparency inset)

Chapter 8:

Table

8.1 Logit analysis of the proportion of highly annoyed by aircraft noise in terms of sensitivity and psychiatric status (from O'Muircheartaigh and Wiggins, 1981)

8.2 Analyses of aircraft noise survey (ANS) data: annoyance (0,1) as dependent variable

8.3 The effect of including interviewer characteristics in the fixed part of the model

- 8.4 Analysis of data from the physically handicapped survey: functional limitation profile (FLP) as a dependent variable for wave one
- 8.5 Two level analysis for PHS wave one with two interviewer level variables in the fixed part of the model with age of respondent as a random effect
- 8.6 Two level analysis of PHS data for separate waves. FLP as the dependent variable
- 8.7 Combined level analysis of PHS data. FLP as dependent variable
- 8.8 Two level analysis for combined waves declaring interviewer as a nested factor

Figure

- 8.1 An illustration of allowing age of respondent to enter the random part of the model

Chapter 9:

Table

- 9.1 Confidence limits for P (%) when $n = 1000$
(taken from Cochran, 1963)
- 9.2 Responses to three requests in a mailed survey
(taken from Cochran, 1963)
- 9.3 Per cent disabled in each stage of survey (taken from Locker et al., 1981)
- 9.4 Fertility among respondents and nonrespondents by age
- 9.5 Mean income among respondents and nonrespondents by age (both from Thomsen and Siring, 1983)

- 9.6 General Household Survey Annual Response Rates (from Lievesley, 1986)
- 9.7 Nonresponse rates by components in Family Expenditure Surveys since 1967 (from Thomsen and Siring, 1983)
- 9.8 Social and Community Planning Research ad-hoc survey response rates (from Lievesley, 1986)

- 9.9 Framework for the analysis of nonresponse data (from Thomsen and Siring, 1983)
- 9.10 Characteristics on several calls of a sample of randomly selected adults from dwellings in the Detroit Metropolitan area (reproduced from Kish, 1965)
- 9.11 Distribution of interviews by the number of calls needed to achieve them
- 9.12 Distribution of interviews by the number of calls made by the interviewer
- 9.13 Main categories of outcome by the number of calls made by the interviewer
- 9.14 Profile of sample for 1985 SCPR Social Attitudes Survey by the number of calls made (all from Lievesley, 1986)
- 9.15 Comparison of percentages of men obtained at the first and second calls with those expected on the hypothesis of random sampling
- 9.16 Percentages of successes at the second call in an experimental survey (both from Bartholomew, 1963)
- 9.17 Proportion of sample reached by repeated calls

9.18 Average response probabilities of those reached
(P_r) (both from Frankel and Dukta, 1983)

Figure

9.1 Some latent response functions
(from Frankel and Dutka, 1983)

9.2 A probabilistic model for nonresponse
(from Thomsen and Siring, 1983)

Chapter 10:

Table

10.1 Two models of mean cumulated costs with 6
calls on not-at-homes (taken from Kish, 1965)

10.2 Estimated bias and mean square error by number
of calls

10.3 Total costs by number of calls

10.4 Strategies that all cost the same

10.5 Mean square error of the sample mean for the
different strategies in table 10.4

(Tables 10.2 to 10.5 all from Thomsen and Siring,
1961)

10.6 Field, and other, costs for NORC surveys

10.7 Average travel cost and marginal cost to
complete interview by number of calls for NORC
call-back samples

(Tables 10.6 and 10.7 from Sudman, 1961)

10.8 Results for different methods of dealing with
nonresponse

10.9 Resumé of information in table 10.8

10.10 Rank analysis of table 10.8

10.11 Average cost of interviews on clustering type 1

10.12 Average cost of interviews on clustering type 2

10.13 Average cost of interviews on clustering type 3

(Tables 10.8 to 10.13 all from Durbin and Stuart, 1961)

10.14 Relative losses (L) for six models of population weights (U_i); for discrete (L_d) and continuous (L_c) weights; for relative departures (k_i) in the range 1 to K

10.15 Losses L for six models of sample weights $u_i = U_i / k_i$; the departures > 1 from 1 to K represent compensations for undersampling

(Tables 10.14 and 10.15 from Kish, 1976)

Figure

10.1 Distances between sample points of different densities in a rectangular area of 960 square miles (taken from Hansen, Hurwitz and Madow, Vol. 1, 1953)

10.2 Schema to help interpret the interplay of costs and variance

Chapter 11:

Table

11.1 Sample design summary

11.2 Items used for evaluation

Figure

11.1 Strategies for retrospective evaluation

Chapter 12:

None

Chapter 13:

Table

- 13.1 Relative bias of successive calls for six items over four call**
- 13.2 Relative success rates for "appointment" versus "non-appointment" calls**
- 13.3 Relative bias for appointment versus non-appointment calls during recalls beyond the initial call**
- 13.4 Estimates for overall means and proportions used in the evaluation**
- 13.5 An illustration of an evaluation for a single item; number of years of full-time education (no. 45)**
 - Strategy (iv): 2nd call appointments
 - North domain: RDW estimation procedure
- 13.6 Efficiency measures for a single item: number of years of full-time education**
 - Strategy (iv): 2nd call appointments
 - North domain: RDW estimation: Cost model III
- 13.7 Efficiency measures for table 13.6 under two different sample bases**
- 13.8 Efficiency indices for table 13.6 under cost models I and II, North domain, RDW estimation**
- 13.9 An illustration of only going to two calls in the North so as to maintain the original achieved sample of 2416**
- 13.10 Average efficiency score for six items under strategy (i) in the North for cost model II**

Figure

13.1 Plot of efficiency scores for table 13.6

**13.2 An illustration of a simultaneous comparison of
four stopping strategies across three cost
models in the Northern domain.**

Chapter 1:

Introduction:

Contents:

1.1: Overview and essential definitions

1.2: Context for evaluation and chapter outline

1.1: Overview and essential definitions

The final outcome of the survey process is typically a survey report containing a set of statements and summary measures which have been generated by a collection of interrelated decisions. These decisions can be conveniently summarised as belonging to four stages.

- i) the conception of a survey; this stage defines the target or ideal parameters, the nature of the survey and its population.
- ii) sample selection; the outcome of tracing these ideal values through the sampling frame
- iii) data collection; the chosen method of data collection translates the ideal survey variables into statistical observations.
- iv) inference; this stage represents a conceptual population of units that would have become available after the data collection and processing stage under a specific set of survey conditions. The success of any survey operation can be measured by the degree of closeness between the target population (stage i) and the inference population (stage iv). Any shortfall between these stages is generated by the implementation of the intervening stages. Differences arise from sampling and non-sampling errors. The magnitude of sampling errors, assuming each individual selected has a calculable chance of inclusion in the sample, can be estimated from the sample itself. Non sampling errors arise from inadequate sampling frames (non-coverage), failure to obtain a response (non-response), poorly formulated questions, incorrect response (response error), and from unintended effects of interviewers or coders on the process of collecting and recording information (non-sampling variance).

Theoretically, for any survey variable, it is possible to capture any shortfall between these parameters by defining a quantity referred to as "total survey error". Adopting a mathematical formulation developed by Hansen et al. (1953) it is necessary to conceptualise a survey as a single trial (t) under a set of "essential survey conditions" (labelled e.s.c), eg. the type of sponsorship or fieldwork agency employed, so that the difference between the value of an estimator (\bar{y}_{tc}) and its' population mean (μ), based on a complete enumeration of individual "true" values for each individual, equals "total error". It is possible to consider this quantity as the sum of two components, namely response bias and response error, the latter a random fluctuation representing the difference between the estimator and its' expected value over all possible repetitions (trials) of the survey under a given set of e.s.c's.

$$(\bar{y}_{tc} - \mu) = (\bar{Y}_c - \mu) + (\bar{y}_{tc} - \bar{Y}_c) \quad (1.1)$$

thus, total error = bias + random fluctuation for particular survey

To help gain a visual appreciation of the various sources of survey error consider the diagram taken from Kish (1965) in figure 1.1 below.

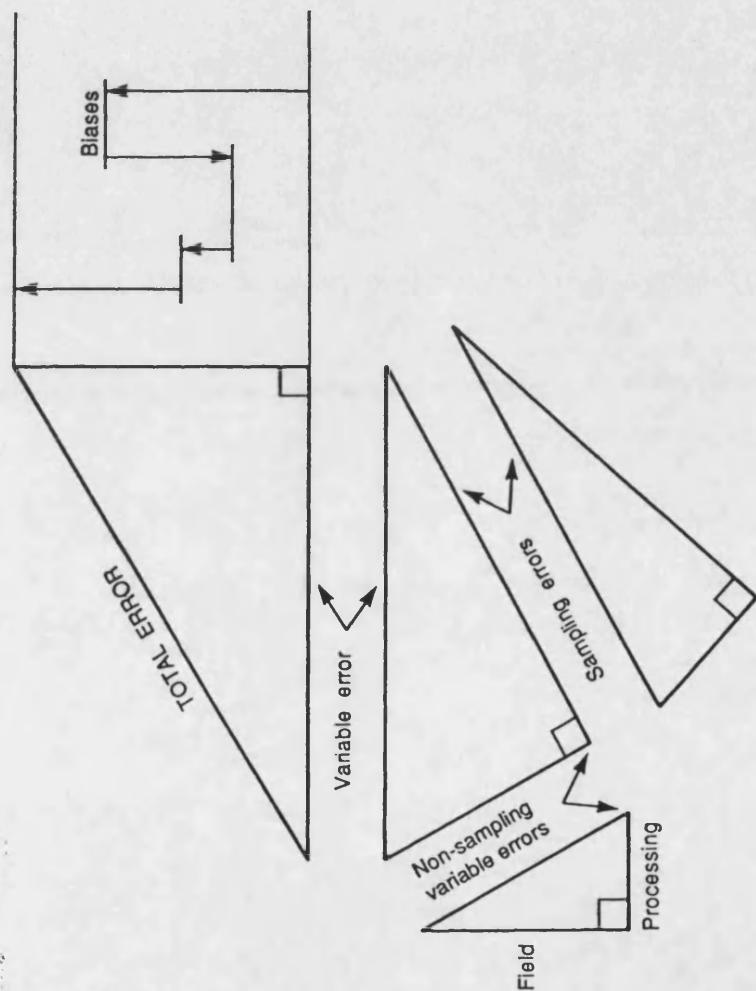
Total error, the hypotenuse on the first triangle is the sum of the uncertainties. Obeying the laws of Pythagoras its' magnitude is simply the square root of the sum of the squared length of the vertical leg, which represents all of the sources of persistent or constant bias, and the squared length of the horizontal leg which represents total variance or variable error. The definition of total error ~~is~~ ^{is} the conventional definition of mean square $\sqrt{\{bias^2 + total variance\}}$

$$\sqrt{\{bias^2 + total variance\}} \quad (1.2)$$

The second term in (1.1) generates an expression for total variance, which can in turn be considered as a sum of separate components each representing a unique source of variable error. Diagrammatically, the horizontal leg in figure 1.1 can be disentangled into a system of sub-triangles representing components of variable error. Sampling error is shown arbitrarily with two components to represent a hypothetical two-stage design.

Conceptually, each aspect of stages ii) and iii) above can eventually find a place in the system of triangles.

Figure 1.1: Total survey error



The thesis focuses on two principal sources of total survey error in interview surveys, namely that arising from the process of observation due to the differential impact interviewers may have on responses and that arising from a major source of non-observation, namely non response bias. The motivation for concentrating effort on these two sources arises from a desire to challenge two common assumptions in survey practice; firstly, that with well trained and supervised interviewers the impact of interviewer effect is negligible and, secondly that it is sufficient to simply report levels of response. In addition, refinement of methodologies to handle these sources of error will facilitate inclusion of other sources of bias or non-sampling error (eg. coder variation) in the expression for total survey error. The specific matter of handling 'sampling error' estimation is assumed to be well documented and catered for both in theory and in practice (see Kish and Frankel, 1976 and Wolter, 1985) and, therefore not a subject for detailed consideration. The inclusion of sample variance estimates in any global appraisal is therefore assumed to be feasible.

Clearly, any global appraisal of the survey process is going to be deficient without paying attention to the costs of obtaining information. Survey budgets are normally fixed, and costed on the assumption that in order to realise the benefits of random sampling more than one attempt will have to be made to secure a response. This results in strict fieldwork norms being established for call-back procedures. Assuming a fixed overall budget a methodology will be established which allows an appraisal of total survey error in terms of maximising accuracy (the inverse of mean square error) for different call-back norms at no additional cost.

In the organisation of the chapters that follow a conventional distinction is maintained between methodologies for handling response errors and those for handling non-response bias. Hence Part One deals with the former and Part Two the latter.

There is also a practical reason for maintaining this division. The empirical evaluations that punctuate the methodology represent historically separate attempts by the author to convince co-researchers of the need to mount methodological investigations without any major increase in the amount of time and effort (costs) involved in the survey operation. This has resulted in a step by step approach rather than a piecemeal one. Altogether the results for three separate empirical studies are presented; two demonstrating methodologies to take account of the impact of

interviewer variability and one to illustrate the impact of non-response bias in the context of a particular call-back and appointments strategy. There is no logical reason why these strategies could not be combined in a single survey if resources permitted.

The illustrations in Part One provide more than a simple response to the requirements of Figure 1.1. The impact of interviewers on response patterns is examined both in a univariate and multivariate context. Secondly, rather than merely accepting responses as the final product of the interview, interviewers enter the analyses as though they were part of the explanatory process. In the context of modelling relationships there will be the opportunity to witness the combined impact of the presence of interviewers and their response/call-back success. This is an attempt by the author to view non-response and call-back planning as a function of the interviewer as much as the response itself.

The final chapter provides a summary of how information on sampling and nonsampling error can be combined with the actual or estimated costs of obtaining data to qualify typical survey analyses and to guide decisions about research design and interviewer work patterns.

1.2: Context for evaluation and chapter outline

As indicated in the previous section Part One uses two studies to illustrate a comprehensive strategy for handling interviewer effect. The work begins in chapter 2 by reviewing the ways in which interviewer behaviour can be described and integrated in a mathematical framework to facilitate empirical evaluation. In order to isolate 'interviewer effect' the researcher will have to modify his/her design to include an element of 'randomisation'. Otherwise, disparities in interviewer response patterns may simply be described to socio-geographic differences between their respondents rather than the interviewers themselves. The degree of randomisation adopted will typically be the product of researcher commitment to such enquiry and the resources available. Evidence of such endeavour is all too rare in the survey literature. Chapter 2 traces the historical use of experimental design to investigate interviewer effect and classifies these studies according to design criteria originally suggested by Hansen, Hurwitz and Madow (1953).

Once the practical demands of randomisation are implemented the researcher then has the responsibility of making effective use of the information generated by the interview. Often, interviewer variability studies conclude with univariate appraisals using 'omnibus' F-tests to flag sensitive items. Such studies serve the author with a stimulus to demonstrate that much more information can be provided. The classical 'analysis of variance' approach becomes a departure point for deeper reflection on the use of the linear model to explore interviewer effect.

This consideration is taken on the assumption that whilst interviewers may be initially assigned equal sized randomly allocated workloads they will rarely complete the same number of interviews (due to nonresponse etc.,). Their achieved workloads render the experimental design as 'unbalanced'.

For this reason the consequence of implementing a particular design will be conducted under the assumption of unbalancedness. Balanced designs are then simply viewed as a 'special' case. Chapter 3 is largely a review of Searle (1971,1987) which facilitates an appreciation of the consequences of using particular experimental designs under the unbalanced data condition. The final outcome of the review is to provide the reader with a guide or pathway as to how s/he can begin to ask basic questions of his/her design. Part of this process, necessarily, requires the researcher to carefully consider the measurement and distributional assumptions underlying any analysis as well as 'signposting' additional possibilities for testing specific hypotheses about interview. The linear model approach allows for the presence of interviewers to be estimated by means of the inclusion of a 'factor' term in the model. There is no reason why this approach cannot be extended to allow the inclusion of other respondent variables in the prediction of a particular survey variable. Such a strategy enables the researcher to see the impact that interviewers may have on the interpretation of any relationships. By considering survey data to be 'hierarchical' (eg. respondents nested within interviewers) the linear model formulation facilitates the inclusion of interviewer level variables (eg. response rates) as well as other survey variables. This permits a more subtle understanding of interviewer variation. These modelling techniques are described in detail in chapter 3.

Chapter 4 extends the theoretical foundations of part one by considering the relationship between univariate summary measures of interviewer effect and those for indices made up of subsets of individual items. The chapter also considers alternative methodologies for univariate appraisal where the design includes replication.

As mentioned above, chapter 5 provides full information about the two empirical studies used to illustrate application of certain aspects of theory covered in chapters 3 and 4. Chapter 6 serves to summarise the main features of the analytical strategies that drive the applications in chapters 7 and 8. Chapter 7 deals entirely with univariate and multivariate assessment of scale scores. Estimates of sampling variability for summary measures of interviewer effect are also included. Chapter 8 concentrates on providing examples of modelling relationships between survey variables in the presence of interviewers. The survey variables included typically refer to scale scores evaluated in chapter 7. Two computational procedures are demonstrated , GLIM (General Linear Interactive Modelling, see Baker and Nelder, 1978) and VARCL (Variance Components software, see Longford, 1988a).

Part Two is structured in a similar fashion to part one. It has as its focus, non-response bias, interviewer call-back procedure and survey costing. The first two chapters provide the theoretical foundation for consideration of non-response bias and costs. The following chapters provide a bridging platform for the illustration that concludes part two. Specifically, chapter 9 reviews the effect of non-response bias and various strategies for exploring the extend of its' impact. In particular, emphasis is placed on response adjustment procedures which take account of call-back policy. eg. Drew and Fuller ,1980. Indeed, chapter 9 completes a survey of methodological material needed to combine the estimates of non-sampling variability covered in part one with estimates of non-response bias. This realises the conceptual demands of Figure 1.1 in the shape of a 'mean square error' estimate for any survey variable. In theory these estimates could be combined with information about costs to enable a retrospective evaluation of call-back policy to be implemented. Unfortunately, the author was unable to obtain actual cost information for the illustration that follows in chapter 13. This obviated the need to review appropriate cost functions that might provide a realistic appraisal and becomes

the main task for chapter 10. Special attention is given to models that assume a call-back/repeated attempt dynamic for obtaining information. The chapter also develops an important 'efficiency' criterion which combines cost and mean square error in a single index for any variable. A suggestion is made as to how the appraisal strategy may be expanded to include multi-item contexts.

Chapter 11 describes the third empirical study to be included in the thesis, namely the Occupational Mobility Survey in Northern Ireland and the Irish Republic (O.M.S.). This study is unusual in that detailed information on the outcome of each call was collected and recorded. Additionally, there was information on the relative success of appointment versus non-appointment calls. This facilitated an evaluation of call-back routes within each wave of calls as well as a straightforward appraisal of the number of calls required to obtain a response. Chapter 12 prepares the reader for the empirical evaluation of the O.M.S study that follows in chapter 13.

The final chapter (14) in the thesis attempts to synthesise the empirical findings in chapters 7,8 and 13 so as to outline the consequences for survey practitioners.

PART ONE

Interviewer variability

Part One: Introduction

Data collected from social surveys are generally obtained from structured interviews with respondents. The data obtained and the conclusions reached depend on the quality of the data collection process. One of the principal sources of error in the data is the method of measurement or observation itself. The form, extent, sources, and effects of such measurement errors are the concern not only of survey design but also of survey analysis.

In the following chapters strategies to examine the differential impact of each of a group of interviewers on the responses in two epidemiological surveys are demonstrated. Although interviewer training attempts to standardise the behaviour of interviewers, differences in style and the resulting interaction between respondent and interviewer traits may result in variations across interviewers for the same respondent. Survey analysts typically assume that these errors cancel out over the sample, leaving sample estimates unbiased.

Theoretically, variation across interviewers will increase the variance of sample estimates. However, to measure the effect requires the randomisation of interviews to interviewers. Not only will this interference increase travel and supervision costs it may endanger interviewer researcher cooperation. Once established we are in a position to examine the impact of interviewer variability in terms of

- a) univariate item analysis
- b) the use of multivariate item sets to construct indices or summary metrics

and

- c) the study of relationships between survey variables

In the following chapter the interview as a vehicle for measurement is contextualised to facilitate an appropriate mathematical foundation for the treatment of interviewer variability, notably via the use of the linear model for the analysis of unbalanced data (chapters 3 and 4). The resulting analytical strategies are then illustrated in the final chapters.

Chapter 2:

The assessment of interviewer effect

Contents:

- 2.1 The interviewer as measurement**
- 2.2 Developing a structure to investigate interviewer effect**
 - (a) The design context**
 - (b) The assessment context**

2.1: The interviewer as measurement

This section provides a basis for the measurement and comparison of interviewer effects; it utilizes Cannell and Kahn (1968) to survey potential sources of effect.

The interview has been defined as conversation with a purpose, the purpose being information "getting". The research interview is a particular form of information gathering labelled measurement. Measurement means assigning of numbers to some population of objects or events, in accordance with some set of rules. The social context in which "the interview" is played out plays a crucial role in this process. Cannell delineates five discrete aspects of the measurement process:

- (i) creating or selecting an interview schedule and a set of rules for using the schedule.
- (ii) conducting the interview (i.e evoking the responses or events to be classified).
- (iii) recording these responses.
- (iv) creating a numerical code.
- (v) coding the responses.

Cannell goes on to suggest that in order to assess and improve the quality of measurement by means of interviewing we need systematic answers to the following questions:

- (i) how can the adequacy of measurement be conceptualized?
- (ii) how can the actual adequacy of any given measurement be determined, in terms of the chosen conceptualization?
- (iii) what can be done to remedy such inadequacies as the preceding steps define and bring to light?

A detailed discussion of the procedures for determining measurement adequacy is considered beyond the scope of this presentation, however, it is important to be aware of three aspects of its consideration, namely, validity, reliability and precision.

Validity is defined as the extent to which an instrument and rules for its use actually measure what they are supposed to. (Sellitz et al, 1959). In particular, do questions look as if they are measuring what they purport to measure? eg. does a set of items which purport to predict psychiatric status actually do so to some acceptable degree? This is often referred to as face validity. (see Cronbach, 1946, Campbell, 1957, Coombs 1964). Contra validity or invalidity has as its hallmark the notion of bias, a systematic or persistent tendency to make errors in the same direction, that is to overstate or underestimate the "true value" of a characteristic. True value is a fundamental in defining bias which may well provoke a misleading sense of simplification. As Cannell adds "there is no such thing as "true value" in the prevailing metatheory of science". External validating criteria may rarely be available in social research. We return to issue of bias in section 2.3 and chapter 3.

Reliability of a measure is defined in terms such as this one: if it (a mode of eliciting responses) is used by different interviewers to measure the same attribute, will it produce the same results? The reader is referred to Cronbach (1960) for further enlightenment.

Precision or sensitivity as related to measurement adequacy concerns the limitations of a measure to discriminate between states of endorsement. For example, if respondents are provided with a forced choice between favouring something or not, we have no idea of the degrees of favour or opposition.

Using an interview to achieve measurement adequacy requires skills to meet the conditions of the interview, both in the creativity of designing the interview schedule and in its field implementation.

What are the conditions necessary for a successful interview? Cannell suggests that even with a reasonable degree of success in attaining uniform validity, reliability and precision interviewers will differ in skill, respondents will differ in motivation and interview content differs in feasibility. He uses three broad concepts to summarize the necessary conditions for a successful interview:

- (i) accessibility of the required information to the respondent.
- (ii) cognition or understanding of the role described for the respondent.
- (iii) motivation of the respondent to take up the role and fulfill its expectancy.

Inaccessibility may arise in a number of ways. Information may have been simply forgotten (Bartlett, 1932), it may be suppressed if painful or embarrassing for the respondent to retrieve or it may be withheld or poorly expressed if the interview conditions fail to stimulate the respondent. This latter point may have a lot to do with socio-cultural aspects and communication.

Cognition requires that the respondent knows and understands what is required and what constitutes successful completion of role requirements. What goes into the development of such understanding depends a lot on the efforts of the interviewer and the sophistication of the respondent.

Motivation is difficult to define as appropriate for a successful interview; Kahn and Cannell (1957) propose a duality, intrinsic motivation, the value attached by the respondent to the interview experience and instrumental motivation; the extent to which the interview is congruent with the respondents own goals.

Research evidence on the interviewing process (Hyman et al., 1954, Rieseman, 1958, Kahn and Cannell, 1957, Richardson et al., 1965) strongly urge that respondent motivation be conceptualized in terms that take account of the social context of the interview. The interview is treated as an interaction between interviewer and respondent, itself being a product of social encounter. One such model is presented in Figure 2.1. Figure 2.2 presents an alternative formulation by Sudman and Bradburn (1974) based on a similar conceptualization.

Fig 2.1 A Motivational model of the interview as a social process

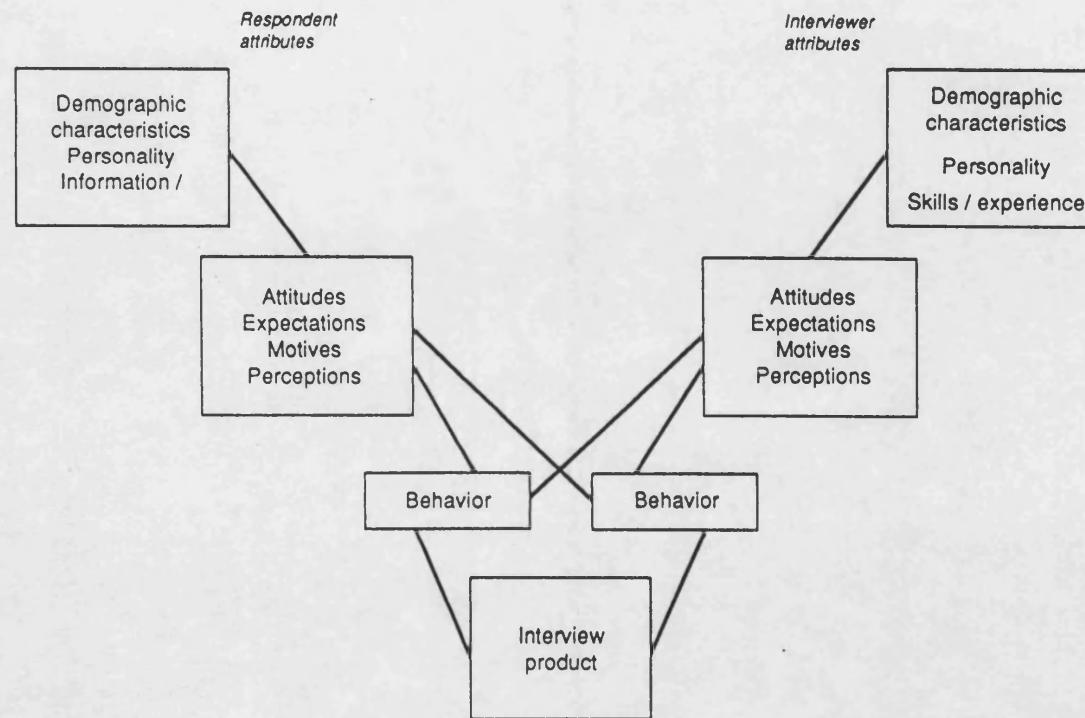
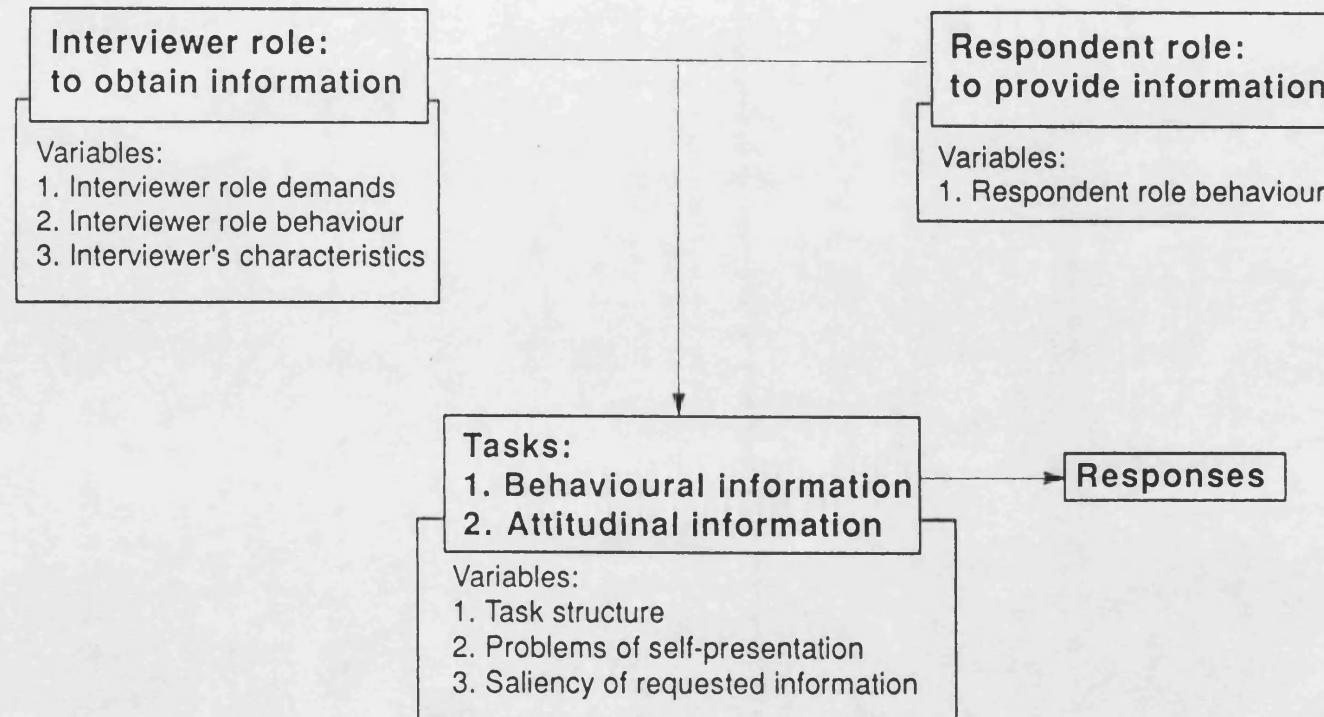
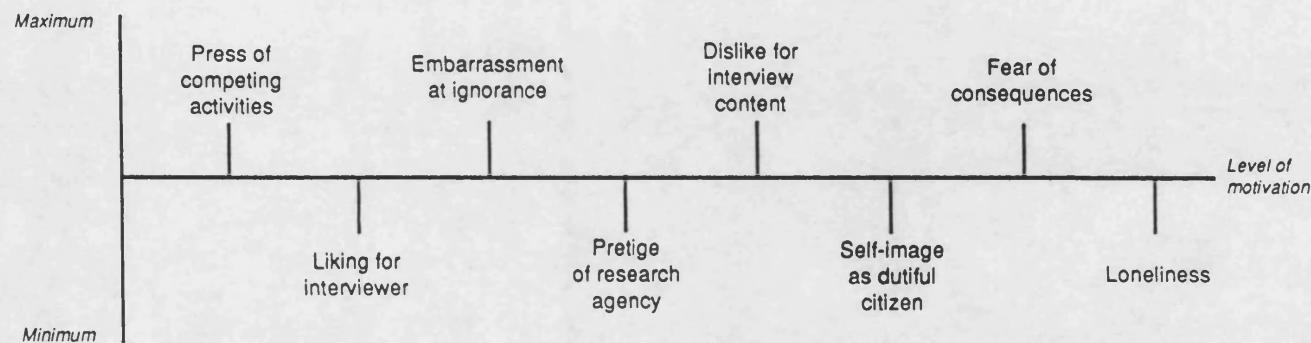


Fig 2.2 Model of interview situation



The models are compatible with the task orientated view of the interview; the interviewer is there to "get" information, the respondent is there to "provide" information. Additional specification of the model, using Lewin (1974) provides insight into the willingness of the respondent to provide complete and accurate data. Here factors identified as opposing forces in Figure 2.3 have been taken from research by Fowler, 1966 and Cannell and Fowler, 1965).

Fig 2.3 Factors affecting the respondent's motivation to provide complete and accurate information to the interviewer.



The initial assessment a potential respondent makes of the interviewer and his/her introductory remarks about the demands, duration, anticipated level of threat and difficulty are critical in determining their transition to take up of the respondent role. Obviously any early commitment or motivational pattern may be dislodged by unexpected interview content or interviewer behaviour. It is therefore useful to consider some of the motivational issues raised by interview content and technique.

Cannell interprets data on respondent motivation in terms of a postulated need on the part of the respondent to maintain self-esteem, so as to be perceived by the interviewer as a "nice" person, someone who does not violate important social norms in thought or act, and to present an image of consistent worthiness! Lamale (1959) analyzing data on the accuracy of consumer expenditure noted that reports were quite accurate for annual expenditure on items like gas and electricity but substantially inaccurate on items like alcohol. Clark and Wallin (1964) describe response bias stemming from reporting on the frequency of sexual intercourse between spouses. Where there was expressed dissatisfaction there was greater discrepancy in report. Parry and Crossley (1950) describe a study to investigate the validity of responses to a set of questions thought to evoke varying amounts of prestige and varying degrees of distortion caused by social pressure, ease of verification and memory factors. On almost every topic significant differences were found between interview data and that provided by records of relevant agencies. The magnitude and direction of the difference suggested a social acceptability ingredient. Cobb and Cannell (1966) report a comparison of two studies, admittedly on very different populations, where respondents were asked to say how willing they were to have their friends know if they had a particular disease. The results are reproduced in table 2.1 below. The rank order correlation between the conditions mentioned by the two samples was perfect for serious conditions and convincing for less serious conditions.

Table 2.1:

Hypothetical willingness to report medical conditions in relation to percentage of cases actually reported

Conditions	Percent willing to report (79 students)	Percent reported* (1388 households)
More serious conditions		
1. Asthma	84%	71%
2. Heart disease	58	60
3. Hernia	55	54
4. Malignant neoplasm	31	33
5. Mental disease	19	25
6. Genito-urinary disease	14	22
Less serious conditions		
1. Sinusitis	89	48
2. Indigestion	88	41
3. Hypertension	83	46
4. Varicose veins	65	42
5. Hemorrhoids	21	38

* Conditions with a frequency of less than 30 percent were excluded as providing unstable estimates of percentage reported.

A study of underreporting of cash loans from loan companies by Lansing, Ginsberg and Bratten (1961) also suggest a social acceptability dimension. The higher income respondents were poor reporters of their loans. Weiss (1968) found that low socioeconomic group mothers were more likely to report that their children repeated a grade in school than are mothers in higher socioeconomic groups. Cannell (1977) suggests that problems of elapsed time, impact, and threat of embarrassment appear to be the most significant issues for research on studies of underreporting.

Clearly cognitive factors help determine the level of reporting accuracy, but it would be mistaken to believe that all such misunderstandings stem from the respondent. Interviewers may often change the wording of a question, omit it altogether so as to suit their own need and sensibilities. (Flowerman et al, 1950).

There is also a powerful interaction between motivation and memory. Problems of motivation would not arise if all information were accessible to respondents. A respondent may have simply forgotten an event (memory decay), for example, Weiss et al (1961) showed that information about jobs and job histories becomes less accurate as time increases. Neter and Wakesberg (1965) found that the date of recent household repairs could be reported accurately, but those occurring more distantly were frequently underreported or misplaced in time (telescoping effect). Another aspect of accessibility has to do with how salient the requested information is to the respondent. Events having strong significance in one's life are usually recalled better than lesser events.

The tendency to suppress information may be related to threatening experience. Janis (1958) found that hospitalized patients could not fully report their preoperative anxieties a few days after surgery.

A fourth type of inaccessibility may result simply because the respondent does not know the information that is required. Reporting accuracy drops when "proxy" respondents are used, Cannell and Fowler (1965).

Interview bias can also occur in terms of the joint effect or interaction of respondent and interviewer characteristics. The encounter is delicate and susceptible to influence. Each person in the interview may have fixed attitudes, personality characteristics and stereotypes of others. Both respondent and interviewer possess visible characteristics which may create false security or hostility. Kahn and Cannell (1957) suggest background characteristics enter the interview in two ways:

- i) as sources of attitude, perception, expectation and motivation. For example a person's gender may determine many of their attitudes to topics such as washing dishes or ironing (Jowell and Airey, 1986).

- ii) as cues in the interaction; for example skin colour may well affect perception of each other. Williams (1964) in a study of black respondents found racial differences between respondent and interviewer only became a potentially biasing factor as perceived social distance became great or interview content threatening.

For the social researcher the underlying objective in evaluating the issues raised so far will probably be the desire to avoid bias and attain valid measures. To the extent that there exist procedures to circumvent the interviewer and his/her potentially biased judgement there are three persuasive procedures. Firstly, the use of probability sampling techniques not least to reduce interviewer decision making in the field selection of respondents. Second, careful interviewer training and finally, diligent questionnaire design. We shall now focus on the last two considerations.

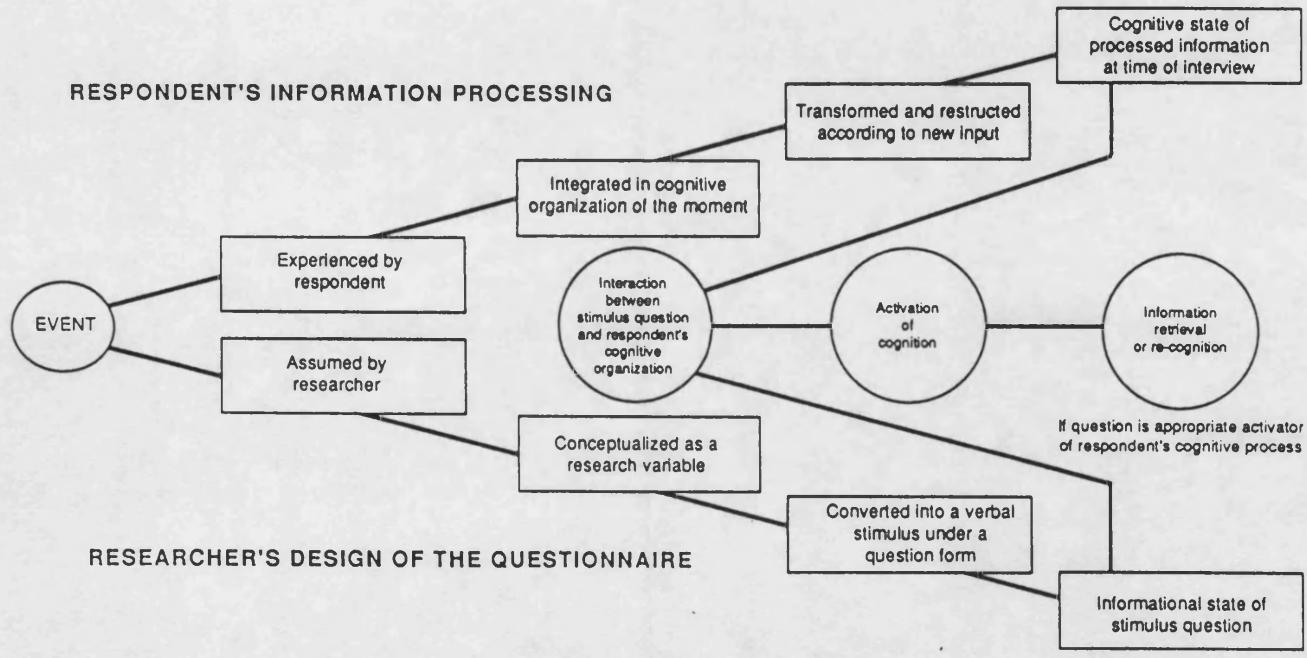
The researcher has to provide questions that can be used verbatim, and the contribute to the conditions of cognition, accessibility and motivation described above. Question wording is vital for respondent cognition. Cannell describes four main cognitive factors in question formulation:

- i) problems of communication; typically a researcher is trying to take into account characteristics of the potential respondents and interview content. Without adequate piloting s/he will remain ignorant of their vocabulary and level of conceptual skill.
- ii) frame of reference; for example inviting someone to respond to the common phrase "How's things?" may invoke ambiguous responses or even highly idiosyncratic interpretation. The respondent will provide their own frame of reference.
- iii) language; the choice of language should be made from a shared vocabulary of respondent and researcher.
A persuasive reason for pre-tests.
- iv) conceptual level of questions; a respondent may understand the language of a question but may find it difficult to respond depending on the degree of familiarity with any implied concepts. For example asking a respondent to respond

to the question "physically disabled persons are just as intelligent as non-disabled persons" may suggest concepts (disabled, intelligent) for which the respondent has only a limited understanding and little experience of having ordered his/her thoughts on the issue.

Where a respondent may never have possessed information, experiences or developed attitudes on a particular subject question wording problems become secondary to issues of inaccessibility. In particular memory and methods to assist recall become pertinent. McGeoch (1932) and interference theory (Postman, 1961) suggest that information does not disappear from memory but may become difficult to recall because of interfering associations. The model presented in Figure 2.4 indicates that the probability of proper recall is a function of the ability of certain stimulus questions to provoke an adequate response from the respondent's cognitive map. The methodological objective becomes one of how to facilitate recall.

Fig 2.4 Model of information processing in the interview. Taken from Cannell, 1977.



The quality of the responses may often be considered solely in terms of the interviewer's technique or ability to motivate the respondent. However as the questionnaire is the basis of the interview careful attention should be given to its design. It is clearly desirable to ensure that the questionnaire does not reduce any opportunity for the interviewer to motivate respondent performance. Respondent fatigue as a direct result of lengthy questionnaires will obviously undermine the quality of the data. Assessing the impact of questionnaire length is difficult and the interested reader is referred to Cannell (1977, p.60). Illuminating discussion on questionnaire structure and question type (open versus closed) is contained in Cannell (1968). We shall now proceed to consider aspects of interviewing technique, selection and training.

It is the interviewer who must bring meaning to the interview experience, make it enjoyable and rewarding. The interviewer has the responsibility of communicating to the respondent their expected role, to let the respondent know how they perform their task. The interviewer may by enhancing respondent motivation or by means of careful explanation render information available that may have otherwise been inaccessible. The primary objective of an interviewer's introductory remarks are to get the respondent "in role"; without implicit cooperation at this crucial stage there may be no interview at all. If successful in securing cooperation an interviewer must ensure that attention is focused on the content of any communication and that the respondent is encouraged to consider each item in accord with the structure of the interview task. Examples of "controlled non-directive probing" (Kahn and Cannell, 1957) have been collected together in table 2.2 below.

TABLE 2.2:

Brief expressions of understanding and interest.

Examples: I see; um-hm; yes, I understand.

Research: Krasner (1958), Quay (1959), Richardson, Hastorf, and Dornbusch (1964), Salzinger (1956), Salzinger and Pisoni (1960), on the ineffectiveness of infrequent encouragement; Mandler and Kaplan (1956), on occasional respondent misinterpretation of "um-hm" encouragements; Hildum and Brown (1956), on the biasing effect of "Good" as an encouragement.

Brief expectant pauses

Research: Gorden (1954) and Saslow et al. (1957), on the positive effects of short pauses (2-3 seconds) as compared to the negative effects of long pauses (in excess of 10-15 seconds).

Neutral requests for additional information

Examples: How do you mean? I'd like to know more of your thinking on that. What do you have in mind there? Is there anything else? Can you tell me more about that?

Research: Guest (1947), Shapiro and Eberhart (1947).

Echo or near repetition of the respondent's words

Example: Respondent - I've taken these treatments for almost six months, and I'm not getting any better. Interviewer - You're not getting better?

Research: No direct evidence, but agreement that sensitive use of the echo conveys close attention, sympathy, and encouragement to continue (Kahn and Cannell, 1957; Richardson, Dohrenwend, and Klein, 1965; Rogers, 1951).

Cannell suggests that little empirical work has been done on criteria for selecting interviewers. Steinkamp (1966) focused on interviewer personality traits as they relate to interviewer effectiveness. Effectiveness was defined in terms of the frequency with which an interviewer obtained information about the holdings reported by financial institutions. It was found that more effective interviewers scored significantly

higher on dominance and intraception¹ tests. They also scored higher in reference evaluations of self-confidence and attention to detail. Interviewers cannot be trained, of course, to modify their demographic characteristics. Freeman and Butler (1976) suggest men and older interviewers tended to demonstrate greater variability than women and younger interviewers. Experience, though whilst not a direct demographic characteristic may be confounded with age, does appear to have a relationship with success in obtaining an interview (e.g. Durbin and Stuart, 1954, Summers and Beck, 1973).

A controversial characteristic of interest is that of skin colour; several authors stress the racial matching of interviewers and respondents yields a pattern of responses different from that obtained in the absence of matching (e.g. Hatchett and Schuman, 1975). Though this effect may only be important when the subject matter is sensitive (Williams, 1964, Boyd and Westfall., 1965).

Most researchers agree that interviewers should be thoroughly trained and briefed in survey procedures, but there is little research as to what methods of training are most effective and/or how much training is desirable. In Cannell's view the most important aspect of a training programme should be the provision of practice and evaluation of interviewing by actually carrying out interviews under controlled conditions. Morton-Williams (1979) provides an interesting discussion of work involving tape recorded interviews as a means of understanding the interview dynamic. Dijkstra (1987) examined the effects of interview style. Using tape recorded interviews and an analysis of coded speech acts, following Brenner (1980), he compared interviewers trained in a "formal" interviewer style with those trained in a "socio-emotional" style. Interviewers in the latter category were found to perform more "person orientated" behaviours and clarifications were often found to be inadequate. Respondents interviewed by the "formal"

¹. Footnote: Intraception indicates an ability to observe others, to understand how they feel about problems and to analyse the motives of others

style interviewers gave more personal information, especially unrelated to the research topic., however, they gave similar amounts of adequate information in response to direct questions. Video taping and feedback would be an obvious extension of such investigation.

Cannell et al (1970) reported that interviewer performance deteriorates over time. Beginning immediately after training and in some cases after a few weeks; one of the main implications of this finding was the need for adequate supervisory and training procedures needed to stimulate or reinforce interviewer role performance during fieldwork. The findings held for both experienced and inexperienced interviewers.

Issues like: who can interview? what is competent interviewing technique? what are the best methods of training that will develop competence? how should interviewer effectiveness be defined? or how should interviewers be supervised? become the active concern of every survey practitioner who is dependent on interviewers to collect information.

However, according to Cannell, they are the mundane regard of research literature. Whilst amenable to experimental evaluation such considerations have only gradually gained methodological momentum. The purpose of this thesis is to begin to redress this imbalance. In particular, attention will be paid to the direction and magnitude of response errors arising from interviewer respondent interactions. The underlying socio-psychological dimension of the interview has been explicitly recognised but in order to transform this dimension into a satisfactory form for empirical evaluation we need a mathematical framework to explore the influence of response errors. Sudman and Bradburn (1974) present a simple model of response effect, defined as a relative measure where,

$$RE = (\text{actual response} - \text{validating response})/ s$$

where "s" is the population standard deviation of the validating information base. Typically in survey interviewers validating information will not be available. It then becomes vital to extend the model formulation. Various mathematical model formulations are reviewed in the following section.

2.2: Developing a structure to investigate interviewer effect

a) The Design Context

When designing surveys to assess the differential impact of interviewers it is necessary to consider methods of selecting the interviewers and assigning them to the various individuals in the sample.

Deming (1960) suggests that the statistical tools required to measure interviewer effect rest on the readiness of each interviewer to work in any area that the random numbers designate for him(sic), at least within a prescribed area. He considers the allocation of two interviewers A and B to work in two possible zones 1 and 2; see table 2.3 (a) below:

TABLE 2.3 (a): **Zone allocations**

Zone	Sample 1	Sample 2
1	A	A
2	B	B

We could never know here whether the difference between the results of A and B could be attributed to the interviewers or to the fact that they worked in different areas. Nor could we be sure that Zone 1 differs from Zone 2: the difference could be attributed to the interviewers. Perhaps more than one apparent sociological or economic difference between 2 areas has given rise to ingenious explanations, when the real difference lay in the interviewers.

The next allocation scheme does not have such a defect, provided the areas where A will work are decided by a random number, and not just because A likes to work there. The difference between the results of A and B may arise from differences between their areas, but not from any difference between the zones, as they will both work in both zones.

Moreover, for the same reason, the difference between the two zones can not be attributed to the interviewers. (The possibility of interaction between an interviewer and an area is a real one, but is beyond the scope of this chapter)

TABLE 2.3 (b): **Zone allocations**

Zone	Sample 1	Sample 2
1	A	B
2	B	A

Taken from Deming (1960)

Now for any survey if we put two interviewers in every two zones and if we let random numbers decide which corners one of them is to work in, we shall be able (a) to compute the variance between interviewers, and (b) to compute the pure sampling error (pure in the sense of being free of the differences between the interviewers).

Clearly under 2.3(a) we could never know whether the difference between the results of A and B could be attributed to the interviews or to the fact that they worked in different areas. Diagram (b) then suggests a minimum requirement for the isolation of interviewer effect, namely a form of interpenetrating sample design which owes its origin to Mahalanobis (1946). Interviews in each zone or primary sampling unit (psu) can be allocated between two (or more) interviewers. Collins (1980) suggests that this minimum design is not difficult or unduly expensive to achieve over at least part of the sample. The examples quoted in his paper (McKenzie (1977), Feldman et al (1951) and Collins (1979) represent the exception rather than the rule in survey literature. Details of such designs are still far from being seen as regular appendages to survey reports in fulfillment of Kish's wishes (1962). Allocation of interviewer workloads on the basis of "nearest to home" whilst ruled out by the requirements of the interpenetrating design still predominate fieldwork practice.

Kish (1962) remarks, research on interviewer variability can be designed to different degrees of symmetry and completeness. A convenient way of summarizing the design context for evaluating the impact of interviews is to adopt the mathematical framework provided by Hansen, Hurwitz and Madow, (1953). The model has five main ingredients, namely:

- i) a population of N individuals and a population of M interviewers, both of which for convenience are assumed to be large
- ii) associated with each individual a true value
- iii) a set of essential survey conditions which determine for a particular individual and interviewer the expected value of a random variable
- iv) zero correlation between the random component of responses for two different individuals with two different interviews

and

- v) the order of interviewing respondents either randomly determined or not affecting the responses

They point out (iv) could be extended, i.e there may be a correlation between responses even when both the individual and the interviewer are different. For example participation in the same training class, common supervision or coding may result in correlated errors for two different interviewers. Fellegi (1964) adopts this approach and includes re-interviewing to estimate components of response variance (see chapter 4).

In the majority of surveys interviewers are available to interview only certain subgroups of the population often in certain geographic areas. Interviewers can then be considered as divided into L groups with M_h interviewers in the h th group who are available to interview a particular N_h individuals and no others. Thus an interpenetrating design, e.g Collins (1979), would be summarized as $L = 32$, $M_h = 2$ and $N_h = 40$. When all interviewers are available to interview all individuals, $L = 1$, $M_h = M$ and $N_h = N$.

Examples of such randomization procedures were seen in Kish's own work, work by O'Muircheartaigh (1976), O'Muircheartaigh and Wiggins (1981) and Wiggins (1985). Other designs found in the literature use varying aspects of experimentation. Examples from the research literature have been classified in this manner in table 2.4 overleaf:

TABLE 2.4:

Classification of major studies of interviewer variability by design schema proposed by Hansen, Hurwitz and Madow (1953).

	<u>No of groups of interviewers</u>	<u>No of Interviewers per group</u>	<u>No of respondents per group</u>
	(L)	(M _h)	(N _h)
One way classification schemes			
Kish (1962)			
Study (a)	1	20	462
Study (b)	1	9	489
O'Muircheartaigh (1976)			
	1	5	130
Collins (1978)	1	19	627
O'Muircheartaigh and Wiggins (1981)	1	8	240
Wiggins (1985)			
Study (a)	1	12	244
Study (b)	1	11	178
Two way classification schemes			
Gray (1956)	1	20	19,28 (two replications by area)
Curtis (1983)	1	2	 (two replications by area)
Nested/interpenetrating designs			
Durbin and Stuart (1954)	3	27,19,119	19,27,4
Gales and Kendall (1957)	24	2	40
Hansen, Hurwitz, Bershad (1961)	125	6	550 (each interviewer completed two assignments per statum)
Kemsley (1965)	24	2	20
Collins (1979)	32	2	20
Curtis (1983)	2	4,3	70,139
Collins and Butcher (1986)	32	2	20

Gales and Kendall (1957) use a sophisticated design involving comparisons of organization (6 levels), briefing (2 levels), questionnaire type (2 levels) and area (4 levels). A completely randomized design would have required 96 pairs of interviewers. Having only 24 available the final design involved four blocks, corresponding to the four areas, and partial replication. Due to one interviewer drop out (not uncommon in practice) the analysis required attendance to missing plots or empty cells. This study, along with Durbin and Stuart (1954) are atypical in the sense that the experimental design (sophistication) was the principal purpose of the survey rather than incidental to it.

Generally factorial designs are uncommon though telephone interviewing makes randomization procedures more readily attainable, even desirable. In addition other sources of non-sampling error, e.g. supervisor or training effect could be routinely included in appraisal of such modes of information collection. There is an attraction that as technology and/or resources permit more complete designs become feasible. However, there is a feeling, echoed in Kish (1962), that it might still be desirable to see more empirical work spread across the breadth of survey research based on modest design protocols rather than a limited number of sophisticated designs in search of definitive truths about interviewer effects. This view is held in the belief that interviewer errors will vary greatly in different contexts. In this sense two level nesting or interpenetrating designs present themselves as the most attractive methodological innovation for the routine consideration of interviewer variability.

(b) the assessment context

By adding further assumptions to their modelling framework Hansen, Hurwitz and Madow facilitate a mathematical assessment of interviewer effect. Firstly, it is assumed n of the N individuals in the population are sampled at random without replacement (extensions to stratified and cluster sampling are also possible). Secondly, m interviewers are selected at random without replacement from the h th interviewer group ($m = \sum_{h=1}^L m_h$) and, finally, an equal number of individuals is assigned to each of the m interviewers. They are considered to be a random subsample of all of the individuals available for interview by this interviewer group. The interviewer workload size is represented by \bar{n} ($= n/m$).

If σ_y^2 represents the total variance of individual responses around the mean of all individual responses in the population. Then following Hansen, Hurwitz and Madow (1953 Vol. 2, Chapter 12)

$$\sigma_y^2 = \sigma_{wy}^2 + \sigma_{by}^2 \quad (2.1)$$

where σ_{wy}^2 is the variance of responses within interviewer groups (taken over all responses of every individual to every interviewer in the group) and σ_{by}^2 is the variance of expected responses for interviewer groups, i.e., between average values for interviewer groups. Now if σ_{yI} represents the covariance between responses obtained by different individuals for the same interviewer, then dividing this covariance by σ_{wy}^2 , we have ρ the intra class correlation between responses of different individuals for the same interviewer. The responses obtained by each interviewer represent "clusters"; the similarity to cluster sampling is apparent if we express $\sigma_{\bar{y}}^2$ as

$$\sigma_{\bar{y}}^2 = (\sigma_{wy}^2 / n) [1 + \rho (n - 1)] + \sigma_{by}^2 / n \quad (2.2)$$

Now, σ_{by}^2 / n represents the variance arising because individuals were sampled independently of the interviewer groups. Having only one interviewer group ($L=1$), implies that interviewer groups serve as strata, so that

$$\sigma_{\bar{y}}^2 = (\sigma_y^2 / n) [1 + \rho (\bar{n} - 1)] \quad (2.3)$$

Hence the effect of using interviewers is to introduce into the variance of y a term involving the intraclass correlation within interviewers assignments. Disregarding this correlation will result in an underestimate of the variance of \bar{y} where there is a substantial interviewer contribution to total variance.

Kish (1962) utilises this approach by expressing interviewer effects as variance components or ρ (rho), an estimate of the proportion of total variance attributable to the interviewers themselves. In the notation of Kish's paper interviewer variance S^2_a is viewed as a component of total variance S^2 , where S^2_b is the variance without any interviewer effect, so $S^2 = S^2_b + S^2_a$, and $\rho = S^2_a / (S^2_a + S^2_b)$. The table summarizing results for ρ is reproduced from Kish (1962) overleaf:

TABLE 2.5

Values of ρ for a number of investigations (reproduced with permission from L. Kish, Journal American Statistical Association, 57,95 (1962)).

		<u>Range of ρ</u>
Kish (1962)		
46 variables in first study	($a = 20$) ¹	0 to 0.07
48 variables in second study	($a = 9$)	0 to 0.05
Percy G. Gray	($a = 20$)	
Eight 'factual' items		0 to 0.02
Perceptions of & attitudes about neighbours' noises		0 to 0.08
Eight items about illness		0 to 0.11
Gales and Kendall ($a = 48$)		
Mostly semi-factual and attitudinal items about TV habits		0 to 0.05
1950 U.S Census ($a = 705$)		
31 'age and sex' items		0 to 0.005
18 simple items		0 to 0.02
35 'difficult' items		0.005 to 0.05
11 'not answered' entries		0.01 to 0.07

¹ a is the number of interviewers in the investigation

Kish's study was an important contribution to the assessment of interviewer variability. He drew attention to the use of traditional analysis of variance techniques to estimate components of variance attributable to the interviewers themselves. It was assumed that interviewers typically completed unequal numbers of achieved interviews. This anchors the study of interviewer variance firmly in the context of the analysis of variance for unbalanced designs. This has important consequences for the following chapter.

Estimates of ρ_{oh} are obtained directly from mean squares in the usual analysis of variance table. Continuing to use Kish's notation.

$$S^2_a = \frac{V_a - V_b}{k} \text{ and } S^2_b = V_b \quad (2.4)$$

where $V_a = \left\{ \begin{array}{l} \text{Mean square for the between interviewers sum of} \\ \text{squares component} \end{array} \right\}$
and $V_b = \left\{ \begin{array}{l} \text{Mean square for within interviewers sum of} \\ \text{squares component} \end{array} \right\}$

and k is shown to be approximately equal to the average workload size, except for a negative correction proportional to the rel-variance of n/m based on work in Anderson & Bancroft (1952, chapter 22)¹.

These design modifications enable the researcher to directly witness the effect of interviewer variability on the precision of estimates. We have seen that for cases when $L=1$ the overall effect of interviewer variance on the precision of a sample mean is to increase it by a factor of $(1 + (k-1) \rho)$. A small value of ρ_{oh} can thus lead to a large multiplier effect. Table 2.6 shows how this effect can be quite dramatic even for moderate sized workloads.

¹ Footnote: the author is aware that Groves, R. is currently conducting research into variation in workload sizes.

TABLE 2.6:

Multiplier effect of interviewer variability on the variance of the sample mean.

Intraclass Correlation Coefficient	Average Workload (k)		
ρ	25	30	35
0.01	1.24	1.29	1.34
0.02	1.48	1.58	1.68
0.03	1.78	1.87	2.02
0.05	2.20	2.45	2.70
0.10	3.40	3.90	4.40

$$\text{Variance of sample mean} = \frac{\text{Variance when no effect present}}{[1+\rho(k-1)]}$$

A value of $\rho_{oh} = 0.3$ and average workload size of 31 implies that if variance estimates are computed without regard to interviewer effect, then the variance will be underestimated by a factor of 2, which will seriously distort any conclusions to be reached from the data.

Univariate assessments of interviewer effect in terms of ρ_{oh} still tend to dominate the few illustrations there are of interviewer appraisal e.g. Collins (1980). Table 2.7 overleaf updates Kish's original review of interviewer variability studies for more recent work.

TABLE 2.7:

Some more results for roh from recent surveys

study	range of roh
O'Muircheartaigh (1976)	
5 factual items	-0.02 to .20
102 attitudinal items	0.00 to .30
Collins (1978, 1979)	
163 items; health survey	
Southampton	-0.03 to 0.05
175 items; consumer survey	-0.02 to 0.05
North Yorkshire	
Groves and Kahn (1979)	
24 items; factual and opinion	
Telephone survey	-0.011 to 0.071
Collins (1980)	
61 items; consumer survey	
Milton Keynes	0.00 to 0.05
O'Muircheartaigh and Wiggins (1981)	
41 psycho-social items	0.00 to 0.09

Groves and Kahn (1979) have applied this method to telephone interviewing research. In addition to using ρ_{oh} as a measure of effect, they define

$$\frac{\text{deff}_{\text{int}}}{\text{SRS wr variance}} = \frac{\text{clustered variance}_{\text{int}}}{(2.5)}$$

where the "clustered variance_{int}" is calculated using clusters defined as groups of respondents interviewed by the same interviewer. The denominator treats responses as though they were a simple random sample. Estimates of ρ_{oh} are obtained using "deff_{int}" ($\rho_{int} = (\text{deff}_{\text{int}})^{-1} / (k-1)$). Since telephone interviewers usually work specific hours during the day they could not make calls on numbers at all hours, and periodically sample numbers were reassigned to interviewers who worked different shifts. What resulted was a randomization within interviewer shifts. What results, according to the authors, are respondent differences across shifts that are largest between those 'reached' between weekday mornings and afternoons on one hand and those reached on weekday evening and weekends, on the other. In the investigation telephone survey interviewers completed an average of 44 interviews. Because the telephone interviewers were able to take on more interviews than their personal survey interviewer counterparts (average workload, c.11) there was a resultant loss in precision for the telephone interviewers. Though this was not always found to be the case. Another interesting speculation is that telephone interviewing may be a larger threat to survey precision than in typical survey interviews where interview workloads are reasonably dispersed. Telephone interviewers work in the same location and there are typically few of them. At the same time, of course, this organization permits more sophistication and routine analysis of the magnitude of interviewer effects.

The use of ρ_{oh} has also been used in the investigation of interviewer effect on questionnaire indices based on multivariate item sets (O'Muircheartaigh, 1976., O'Muircheartaigh and Wiggins, 1981) and exploration of the structure of interviewer effect in the context of multivariate analysis (O'Muircheartaigh, 1977, Wiggins, 1985).

In the latter cases the estimation of an interviewer effect has been based on a fixed effects one way analysis of variance model. Such apparently divergent assumptions in applications have not always been made specific. More details on these multivariate approaches are given in chapter 4.

Kish (1962) also investigates the behaviours of "synthetic" roh for the exploration of interviewer effects on subclass means, where the impact is seen to be smaller in accord with $(1 + p(n^* - 1))$, where n^* is the average number of interviews per subclass per interviewer. Comparisons of subclass means, according to Kish, tend to reflect zero interviewer effect in accord with an additive model of the effects. Surprisingly little empirical investigation has been carried out on subclass analysis. A notable exception being the work by Hansen et al (1961).

The methodology illustrated by Kish's work has had a profound impact on the strategies of methodologists researching this area for the last two and a half decades. In spite of the importance of the subject there still appears to be varying degrees of confusion as to what the basic analytical framework should be, what the various studies achieve and what questions should be asked of the findings. To help establish the full range of practical information available to researchers investigating interviewer effect it is felt desirable to clarify the implications of any underlying assumptions conveyed by the analytical framework. Modest, or even mild sophistication in experimental design for unbalanced data (unequal workload sizes) can provoke difficulties for interpretation. A major objective of this thesis is to present an orderly account of the various approaches and to discuss the practical use of the findings. The following chapter develops this proposition by beginning with an orderly review of the assumptions underlying the model framework presented so far and those inherent in analysis of variance. Whilst statistical theory by itself cannot ask the right questions it will be useful to review the development of analysis of variance methods to facilitate a clearer understanding of the implications of "modelling practice" for the investigation of interviewer effect.

Chapter 3:

A structure to investigate interviewer effect: the linear model

Contents:

Introduction

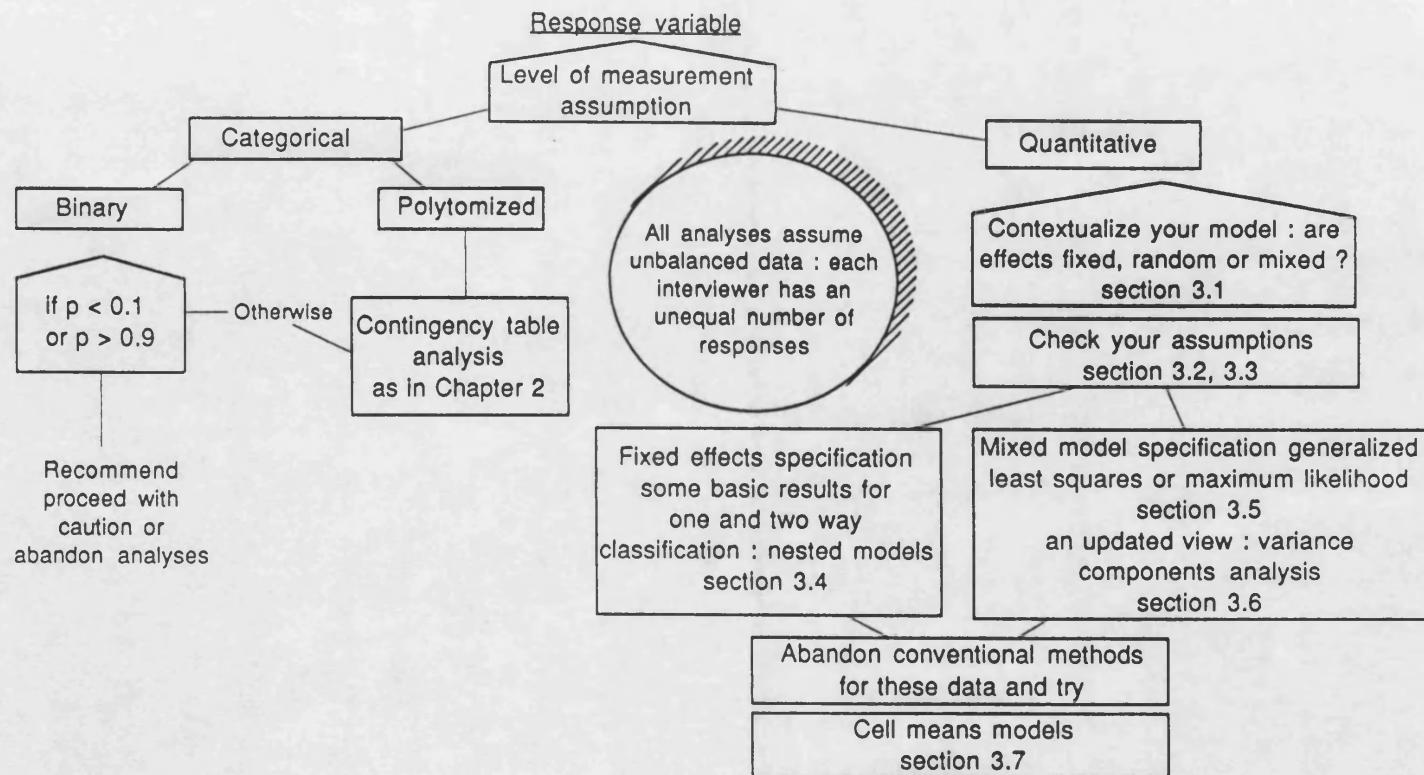
- 3.1 Departure from basic assumptions**
- 3.2 Fixed or random?**
- 3.3 Assumptions underlying fixed and random
effects models**
- 3.4 The linear model under the fixed effect
assumption: a review**
- 3.5 What happens to the linear model if some or all
of the effects are random?**
- 3.6 Maximum likelihood estimation**
- 3.7 A pathway for the exploration of interviewer
effect**

Introduction:

The necessary design modifications for the investigation of interviewer variability have important consequences for the analyst. To gain a proper understanding of these consequences it is useful to examine the founding principles of analysis of variance. One important practical consideration is that whereas interviewers may begin their fieldwork commitment with equal allocations of work, they rarely complete the same number of interviews. It is, therefore, appropriate to review analysis of variance procedures in the context of unbalancedness or unequal achieved workloads. Equal achieved workloads is then, simply, a special case of the more general situation of unbalancedness. This view neatly accords with Searle (1971, 1987), and thus the chapter is largely a review of his work¹. The object is to present the reader with a 'guide' to the implications arising from the introduction of any experimental considerations so as to make proper use of his/her data. We begin with a summary in figure 3.1 of the structure of the chapter. The concluding section, 3.7, presents a fuller version of this scheme by including the main findings arising from the intervening review sections.

¹Footnote: Throughout the chapter specific page and section numbers refer to Searle (1971).

Figure 3.1: A guide to the way theory can help determine strategies to investigate Interviewer effects.



3.1:

Departures from basic assumptions

(i) **measurement level of response variable, y:** survey questionnaire items are typically polytomized quantities, sometimes binary or often reduced to binary categories for analysis. The typical measure for univariate assessment of interviewer effect is ρ , the intra class correlation coefficient, based on a variance component model which assumes continuity for y . Collins (1980) and Anderson and Aitken (1985) suggest as a "rule of thumb" that applications of analysis of variance to situations where y is binary are reasonably accurate so long as the proportions in each of the response categories are between 0.1 and 0.9, i.e. values smaller than 0.1 or larger than 0.9 seriously undermine the assumption of constant variance. Cox (1970) suggests that treating binary observations just as if they were quantitative observations is reasonable in the range 0.2 to 0.8, and within this range there is unlikely to be any serious loss in efficiency arising from changes in the variance. There appears to be no definitive position on this question, the reader is left to consider the behavior of the variance as proportions change within the range 0 to 1 in table 3.1 below. Beyond preliminary inspection of the overall proportion of respondents endorsing particular categories prior to calculating summary measures of effect as an alternative to traditional analysis of variance it is possible to treat any binary item as a 0/1 response in the context of logistic regression with the interviewer as a factor (see Cox (1970) and Wiggins and O'Muircheartaigh (1981). More recently via Anderson and Aitken (1985), Pannekoek (1988, 1989) and Wiggins, Longford and O'Muircheartaigh (1990), we witness encouraging use of appropriate models for non-normal data in interviewer variability appraisals.

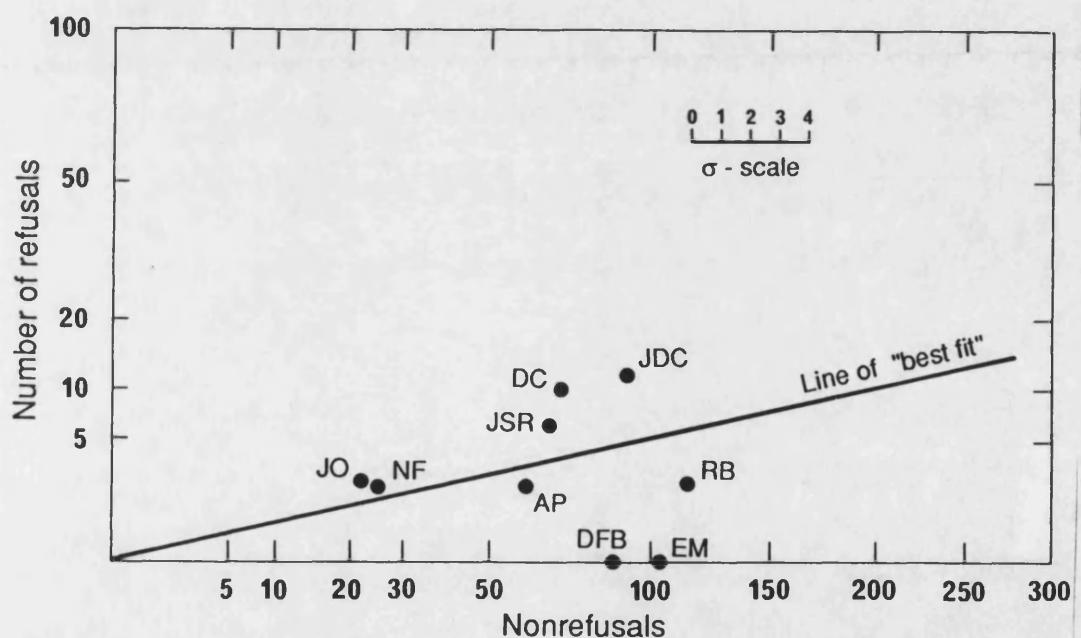
Table 3.1: The Curves of Binomial Variation.

$\bar{Y} - P$	0.001	0.005	0.01	0.05	0.10	0.20	0.30	0.50	0.70	0.80	0.90
$\sigma' = P(1 - P)$	0.001	0.005	0.010	0.048	0.09	0.16	0.21	0.25	0.21	0.16	0.09
$\sigma = \sqrt{P(1 - P)}$	0.03	0.07	0.10	0.22	0.30	0.40	0.46	0.50	0.46	0.40	0.30
$\sigma' / \bar{Y}' = (1 - P)/P$	999	199	99	19	9.0	4.0	2.3	1.0	0.42	0.25	0.11
$\sigma' / \bar{Y} = \sqrt{(1 - P)/P}$	31.6	14.1	9.9	4.4	3.0	2.0	1.5	1.0	0.66	0.50	0.33

(Taken from Kish, 1965)

Deming (chap. 13, 1960) provides a useful graphical technique for evaluating the performance of interviewers on dichotomous outcomes (both on individual items (plots of "Yes" versus "no" responses) and as indicators of interviewer performance eg. plots of refusals versus non refusals). Figure 3.2 is taken from Deming to illustrate the idea.

Figure 3.2: Refusals and nonrefusals for 9 interviewers at the end of 4 weeks. Interviewers DFB and EM are significantly superior, or else there is something wrong with the records.



(Taken from Deming, 1961).

Rather than simply treating polytimized responses as continuous it may be possible to devise alternative measures of interviewer variability. Indeed Gales and Kendall (1957) suggest the statistic $V = \chi^2 / k-1$ in a situation where interviewers are comparable in pairs and the number of respondents falling into the response categories of each interviewer workload are arranged in a 2^*k table. If the interviewers are in complete agreement in the sample then $V = 0$. As the rows (or interviewer response patterns) become more divergent V increases without limit. Clearly this approach can be extended to handle more interviewers as rows in the table and alternative contingency coefficients may be utilized to summarize interviewer variability (refer Everitt, 1977). Another approach might be to analyse response patterns for individual items as m (number of interviewers) * k (response categories) tables by means of log linear modelling (O'Muircheartaigh and Payne, Chap. 4, Vol. II). Use of conditional odds ratios or the mean deviance associated with the independence model might serve as useful indicators of interviewer variability. Essentially the independence model coincides with the assumptions made in applying a conventional chi-square analysis and thus assumes no association between the influence of an interviewer and the response category in which an individual respondent places her/himself. Large mean deviance would indicate a poor fit for such an assumption.

(ii) **equality of variance:** even where quantitative measurement can be safely assumed for y there is no guarantee of the assumption of "equality of variance". Conventionally it is recommended that this be the first assumption to be subjected to a statistical test. Scheffe (1959) indicates that the standard test for equality of variance (Bartlett's 1937 test) tends to mask differences when they exist if the kurtosis, $k < 0$ and to find differences when none exist when $k > 0$; for some populations with $k > 0$ the test is sensitive to non-normality. If the variances are equal but the data are non-normal with $k > 0$, the preliminary test is then likely to reject the hypothesis of equality of variance and the user will accordingly refrain from applying analysis of variance where it may be appropriate. Though, Scheffe adds, where one is in a situation with balanced designs reasonable protection is afforded against these reservations. This comes as small comfort, since rarely are investigations of interviewer variability carried out under the 'balanced' condition.

As a quick approximate analysis in the case of unbalanced data, or unequal cell numbers, Scheffe recommends an analysis of cell means (equivalent to a layout with one observation per cell). If the results are sufficiently conclusive it may suggest that the tedious calculations necessary for the unbalanced case may be unnecessary.

(iii) normality: Scheffe provides further comfort for the investigator in that non-normality has little effect on inferences about means (even extending to certain methods of multiple comparison, the S-method in his text) but has serious effects on inferences about variances of random effects whose kurtosis differs from zero. He suggests that ordinarily we don't have any idea of the magnitude of the kurtosis of the effects measured by the variance component other than the error component. He suggests "The situation is not very hopeful, and normal theory inferences about variance components must be accepted as being much less reliable than those about means. The conclusions reinforced by consideration that models with variance components have, even without the normality assumption, a rather tenuous relation to those frequent applications where nothing is done to insure the random sampling of effects which is assumed in the model" (see section 3.2).

(iv) independence: the effect of correlation in the observations are formidable to cope with. Scheffe presents the case for a two-way layout with one observation per cell in which observations within a column are serially related but that the columns of observations are independent of each other. Then it is possible to estimate the impact on inferences. It is difficult to envisage a similar situation arising in the context of interviewer variability unless pairs of interviewers were re-interviewing over successively small time intervals.

Transformations, are commonly suggested as ways of reducing non-normality or more commonly to reduce inequality in variance. Many of these transformations are special cases or modifications of a general transformation proposed by Freeman and Tukey (1950) where the standard deviation of a random variable can be expressed as a function of its mean, ie where $\sigma_y = \phi(\mu)$. For example the binomial distribution of a proportion of yes's has the required properties

$$E(y) = np, \sigma(y) = [np(1-p)]^{1/2} \text{ so that } \phi(\mu) = [\mu(1-n^{-1}\mu)]^{1/2} \quad (3.1)$$

The object then is to find a transformation $z = f(y)$ so that the standard deviation of z is at least approximately constant. In the case of the binomial this results in an "angular transformation" (see p. 365).

Another possibility might be to simply take logarithms. However it is important to remember that transformations transform the mean as well as the variance. As Scheffe points out in testing the hypothesis of equal group means in the one way layout, transformation would seem to cause no difficulty since a "1:1" transformation of the original means are equal if and only if their transforms are equal. But if after a transformation the 'equal means' hypothesis is rejected by an F-test and a multiple comparison test is desired, the original scale of means may make more sense than the meaningfulness of comparison or estimation on different scales.

Section 3.2:

Fixed or random?

A suitable model for the study of interviewer effects when we have

$L = 1$, $M_h = M$, and $N_h = N$ under the Hansen, Hurwitz and Madow design parametrization (see chapter 2) is the one way classification model

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (3.2)$$

where y_{ij} is the recorded value of y for the j -th individual in the workload of interviewer i , μ denotes the overall mean of the observations, α_i being the effect or "net bias" of the i -th interviewer. This equation would be the same whether we had assumed that the interviewers had been a random sample from a large population of potential interviewers or if our interest had focused solely on the group of interviewers selected for the experiment. The formulations are different in the two instances because of the interpretation attributed to the effects. In the first case, the typical one implicit in many empirical reviews, interviewer effects are assumed to be random, in the latter fixed.

The decision as to whether effects are fixed or random may not always be obvious. For instance it may not always be realistic to assume the existence of a large pool of available interviewers.

As Searle (1971) suggests when endeavouring to decide whether a set of effects is fixed or random, the context of the data, the manner in which they were gathered and the environment from which they came are the determining factors. Perhaps by focussing on the outcome of such a decision an appreciation may be clearer: are inferences going to be drawn from these data just about the levels of the factor alone? "Yes" - then the effects are to be considered as fixed. "No" - then, presumably, inferences will be made not just about the levels occurring in the data but about some population of levels of the factor from which those in the data are presumed to have come.

In the author's view the control an investigator has over the random selection of interviewers for inclusion in an experiment may not be direct. In large agencies fieldwork managers may preselect experimental groups on the basis of willingness to participate in such experiments or commitment to the agency. It would therefore seem pragmatic to regard interviewer effects as fixed unless there are convincing grounds to believe otherwise.

The assumption of randomness of selection of interviewers does not automatically carry with it the assumption of normality. This assumption is often made for random effects, indeed the calculation of \hat{p} is based on the F statistic obtained in analysis of variance where

$\hat{p} = F - 1 / (F - 1 + k)$ and quoted significance levels are based on an evaluation of the F ratio. The majority of estimation procedures for variance components do not require normality unless the distributional properties of resulting estimators are to be investigated.

Section 3.3:

Assumptions underlying fixed and random effects models

The fixed effects model, when all the terms are fixed apart from the error term, was named model I by Eisenhart (1947) and the random effects model, model II. In each model error terms are a random sample from a population distribution as $(0, \sigma_e^2 I)$. In the random model the α 's are also a random sample from a population distribution as $(0, \sigma_\alpha^2 I)$, whereas in the fixed model the α 's represent fixed or specific effects. Furthermore in the random model sampling of the α 's is assumed to be independent of that of the e 's so covariance between the α 's and the e 's is zero.

The words of Eisenhart still have important currency today: "The failure of most of the literature on the distinction between problems of class I (i.e. model I) and class II (model II) is very likely due to two facts: first, the literature of analysis of variance deals largely with tests of significance in contrast to problems of estimation; second, when the analysis of variance is used merely to determine whether to infer (a) the existence of fixed differences among the true means of the subsets concerned or (b) the existence of a component of variance ascribable to a particular factor, the computational procedure and the mechanics of the statistical tests of significance are the same in either case - the same test criterion (F or z) is evaluated and referred to the same levels of significance in either case. On the other hand, in the estimation of the relevant parameters, and in the evaluation of the efficiency or resolving power of a particular experimental design, the distinction between these two classes of problems needs to be taken into account, since in the problems of class I the parameters involved are means and the issues of interest are concerned with the interrelations of these means, i.e., with differences between pairs of them, with their functional dependence on some independent variable(s), etc.; whereas in problems of class II the parameters involved are variances and their absolute and relative magnitudes are of primary importance." Basically the mathematical models appropriate to the two classes differ, and so do the questions to be answered by the data.

It is also important to remember that the formula and procedures of analysis of variance are used to merely summarize properties of the data in hand, no assumptions are needed to validate them. On the other hand when analysis of variance is used as a method of statistical inference, then certain assumptions about the "population" and sampling procedure by means of which the data were obtained, must be fulfilled if the inferences are to be valid.

Thus under model I the parameters are population means. The y_{ij} are observed values of random variables distributed about true mean values μ_i ($i = 1..m$) that are fixed constants. These means are expressed as a simple additive function of the corresponding marginal mean and the general mean, that is

$$\mu_i = \mu + (\mu_i - \mu) \quad (3.3)$$

When the assumption of randomness and additivity is satisfied then the difference between any two interviewer means is an unbiased estimator of the general average difference of the interviewers concerned. Thirdly, the random variables y_{ij} are homoscedastic and mutually uncorrelated, i.e. they have a variance-covariance structure as stated earlier $(0, \sigma_e^2 I)$. As Eisenhart indicates, if the first assumption is satisfied but if either or both of the latter two are not, then the strict validity of analysis of variance vanishes out of the window. Even if all three assumptions apply it is not possible to conduct exact tests of significance based on the y_{ij} alone. Normality of the joint distribution of the y_{ij} in addition to those already mentioned make it possible to apply all of the usual analysis of variance procedures for estimating, and testing to determine whether to infer the existence of, fixed linear relations, e.g. non- zero differences among the population means.

Under Model II the parameters are components of variance. By following a line of reasoning similar to the one for model I three basic assumptions are necessary but not sufficient for the use of exact tests of significance. Namely, random variables, the observed values y_{ij} are now considered to be distributed about a common mean μ , where μ is some fixed constant; additivity, the random variables y are sums of component random variables

$$y_{ij} = (\mu_i - \mu) + \mu + e_{ij} \quad (3.4)$$

(now α_i)

where the $(\mu_i - \mu)$ and e_{ij} are random variables; these random variables are distributed $(0, \sigma_\alpha^2 I)$ and $(0, \sigma_e^2 I)$. The first assumption involved in describing the model brings the problem within the province of mathematical statistics, the second brings meaning to the concept "components of variance" and the third renders each component of variance assignable to a specific factor. Finally assuming all deviations $(\mu_i - \mu)$ and e_{ij} to be normally distributed determines the possibility of testing whether the existence of components of variance is strictly valid.

3.4:

The linear model under the fixed effect assumption

General linear models consist of a model equation with allied assumptions. Historically, they were introduced to explain Fisher's analysis of variance approach and associated estimation. More recently the danger is that they have been elevated to the role of completely describing all of the statistically interesting features of an "experiment", thereby, exerting substantial effect on the analysis itself. Confusion is often fostered by statements like "sum effects zero" or "functions estimable". The objective of the presentation that follows is to provide the potential investigator of interviewer effects with

(a) a clarification of the analysis of variance procedures appropriate to the underlying experimental design, and

(b) to eliminate any possible confusion by keeping in mind basic fundamental assumptions about how a model relates to its associated context.

The context will always be assumed to be 'unbalancedness'. Basic results for one, two way classifications and nested designs will be presented under the unifying theory of the linear model. 'Balanced' designs are simply a special case of each description.

The author is greatly indebted to the work and inspiration of S. R. Searle. As we have seen in the previous section under the fixed effect assumption we are concerned with making conclusions or inferences that are confined to the interviewers actually studied specific to a particular time and survey. We are interested in detecting and estimating fixed (or constant) relations among the interviewers.

Writing the model ($y_{ij} = \mu + \alpha_i + e_{ij}$) specified in section 3.1 in matrix form we have

$$y = Xb + e \quad (3.5)$$

where y is a vector of $N \times 1$ observations
 b is a vector of $p \times 1$ parameters
 X is a matrix of known values, in most cases 0's or 1's
 e is a vector of random error terms such that $E(e) = 0$ and $E(y) = Xb$

The assumption $e \sim (0, \sigma_e^2 I)$ mentioned in section 3.3 is the only one necessary for point estimation, whereas for hypothesis testing and confidence intervals we assume normality of errors.

The 'normal' equations corresponding to the model are:

$$X'X b = X'y \quad (3.6)$$

$$(pxp) (px1) \quad (pxn)(nx1)$$

Unlike regression the matrix X is not full rank so the procedure for solving these equations is to find the generalized inverse of $X'X$ such that

$$b^o = G X'y \quad (3.7)$$

The symbol b^o is introduced because here the equations have no single solution for b . Strictly b^o should be referred to as a solution not as an estimator. $X'X$ is singular so there are infinitely many solutions. Use of the generalized inverse G is a means of skirting the problem of a model "not of full rank". Searle sets out a procedure for a solution based on ascertaining the rank of $X'X$. If $X'X$ has order p and rank r then the user sets $(p - r)$ elements of b^o to zero, striking out the corresponding normal equations leaving a set of r equations of full rank (see Searle, (1971), 5.7(c)). Whilst it is easy to appreciate how this procedure leads to a solution is not immediately obvious for specific applications which $(p - r)$ elements to set to zero. However, having a procedure to derive b^o enables us to formally introduce analysis of variance. Having obtained a value for b^o the expected value of y corresponding to its observed value is $y = Xb^o$ (or $XG X'y$) and the residual sum of squares is

$$SSE = \mathbf{y}'\mathbf{y} - \mathbf{b}^0\mathbf{X}'\mathbf{y} \quad (3.8)$$

the sum of squares due to fitting the mean is

$$SSM = N\bar{y}^2 \quad (3.9)$$

the sum of squares of the model is

$$SSR = \mathbf{b}^0\mathbf{X}'\mathbf{y} \quad (3.10)$$

and the total sum of squares is

$$SST = \mathbf{y}'\mathbf{y} = \sum y_i^2 \quad (3.11)$$

where $\sum y_i^2$ represents the sum of squares of the individual observations.

Hence $SSE = SST - SSR \quad (3.12)$

Correcting for the mean we get

$$SSR_m = SSR - SSM \quad (3.13)$$

$$SST_m = SST - SSM \quad (3.14)$$

and the coefficient of determination

$$R^2 = SSR_m / SST_m \quad (3.15)$$

These partitions of sums of squares form the basis of traditional analysis of variance tables.

On the basis of normality for the error terms we obtain

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I})$$

and $\mathbf{b}^0 \sim N(\mathbf{H}\mathbf{b}, \mathbf{G}'\mathbf{X}\mathbf{G}\sigma^2)$: since \mathbf{b}^0 is a linear function of \mathbf{y}

and \mathbf{H} is defined as $\mathbf{G}'\mathbf{X}$

$$\begin{aligned}
 \text{Formal hypothesis testing is now possible. A test based on} & \quad F(R) \\
 = \text{SSR/a} & \\
 \hline
 & \text{SSE/(N-a)}
 \end{aligned}$$

(where $a = \text{rank } (X)$) cannot be described as testing $H:b = 0$, because b is not what is referred to as 'an estimable function'. Essentially, in our context this means "all interviewer effects zero "cannot be tested. However, certain functions of b can be tested, where $F(R)$ is the appropriate statistic, as discussed below.

Thus we have little use for b^0 as it stands so what about linear functions of b^0 ? Principally, the only ones which are of any interest are those which are invariant to whatever solution it is obtained for the normal equations. Functions such as these are known as estimable functions: basically, they are a linear function of parameters for which an estimator can be found from b^0 that is invariant to whatever solution of normal equations is used for b^0 . A linear function is said to be estimable if it is identical to some linear function of expected values of y . This means that $q'b = t'E(y)$ for some vector t' .

The only hypotheses that can be tested are ones which involve estimable functions. For a one way classification, under the definition of an estimable function consider a t which has one element equal to unity and the others zero; $t'E(y)$ will be estimable and it will be an element of y , i.e. the expected value of an observation.

Thus $E(y_{1j}) = \mu + \alpha_1$ and $E(y_{2k}) = \mu + \alpha_2$

Hence $E(y_{1j} - y_{2k}) = \alpha_1 - \alpha_2$ and therefore $\alpha_1 - \alpha_2$ is an estimable function, alternatively any linear combination of estimable functions is estimable. Furthermore the value of t' is not as important as the notion of the existence of some t' . Whenever $q'b$ is estimable, $q'b$ is invariant to whatever solution of $X'Xb = X'y$ is used. Finally, the best linear unbiased estimate ("b.l.u.e.") of $q'b$ is $q'b^0$, written $q'b$.

These fundamental properties of estimable functions are explicitly presented in Searle (p. 181). Applying them to the one way classification we have the expected value of any observation as estimable: thus $\mu + \alpha_i$ is estimable and correspondingly the b.l.u.e. of $\mu + \alpha_i$ is \bar{y}_i . Using Searle's technique for solving the normal equations we get

$$\widehat{\mu + \alpha_i} = \mu^0 + \alpha_i^0 = \bar{y}_i. \quad (3.16)$$

This is a basic result which provides the b.l.u.e.'s of all other estimable functions. In particular, for scalars λ_i

$$\sum_{i=1}^a \lambda_i (\mu + \alpha_i) \text{ is estimable, with b.l.u.e. } \sum_{i=1}^a \lambda_i \bar{y}_i. \quad (3.17)$$

It is clear from this formulation that the variance of the b.l.u.e. depends solely on the variances and covariances of the \bar{y}_i , namely

$$\text{var}(\bar{y}_i) = \sigma^2 / n_i \quad \text{and} \quad \text{cov}(\bar{y}_i, \bar{y}_k) = 0 \text{ for } i \neq k$$

From this confidence intervals on $\sum \lambda_i (\mu + \alpha_i)$ can be obtained. Rewriting $\sum \lambda_i (\mu + \alpha_i)$ as $\mu \sum \lambda_i + \sum \lambda_i \alpha_i$ (3.18)

enables certain implications to be appreciated. Note, as earlier, $r(X) = a$, so using Searle (section 5.4f) the maximum number of LIN estimable functions is a . All other estimable functions are of the form above. Consequently specific results for the one way classification follow:

Individual terms μ and α_i are not by themselves estimable. For example if we wished to estimate individual interviewer effects then in the second term of (3.18) we must have $\lambda_k = 1$ and $\lambda_i = 0$ for all $i \neq k$. But, then (3.18) becomes $\mu + \alpha_k$. Hence α_k is not estimable. However a simple restatement of (3.17) as $(\sum \lambda_i) \mu + \sum \lambda_i \alpha_k$ estimable for any λ_i made for the purpose of emphasising the estimability of any linear combination of the μ and the α 's in which the coefficient of μ is the sum of the coefficient of α provides us with two functions of particular interest namely:

$$\mu + \sum \frac{n_i \alpha_i}{N} \text{ with b.l.u.e. } \bar{y}_i \text{ i.e. } \lambda_i = n_i / N \quad (3.19)$$

$$\text{and } \mu + \frac{1}{a} \sum \alpha_i \text{ with b.l.u.e. } \frac{1}{a} \sum \bar{y}_i \text{ i.e. } \lambda_i = 1 / a \quad (3.20)$$

For balanced data $n = n_i$ for all i and above expressions are the same. Two special case results also follow: putting $\lambda_i = 1$ and $\lambda_k = -1$ and all other λ 's zero shows $\alpha_i - \alpha_k$ is estimable for every $i \neq k$ i.e. the difference in the net effect between any pair of interviewers is estimable, together with corresponding confidence intervals based on normal theory. Also $\sum \lambda_i \alpha_i$ for $\sum \lambda_i = 0$ is estimable. Thus the linear combination of any of the effects is demonstrated where the sum of the coefficients is zero, e.g. consider that three of our eight interviewers are male (the first three subscripts for convenience) then

$5\alpha_1 + 5\alpha_2 + 5\alpha_3 - 3(\alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 + \alpha_8)$ is an example of an estimable function which actually tests for a sex difference in the resultant effects with b.l.u.e.

$$5\bar{y}_1 + 5\bar{y}_2 + 5\bar{y}_3 - 3(\bar{y}_4 + \bar{y}_5 + \bar{y}_6 + \bar{y}_7 + \bar{y}_8)$$

If a hypothesis of the form $H : K'b = m$ is to be tested, then results from the full rank case (see Searle section 3.6) suggest that $K'b - m$ will be part of the test statistic which, of course, will need to be invariant to b . It will be invariant only if $K'b$ is estimable. Consider the case in a one way classification, where we wish to test $H: E(\bar{y}) = 0$; this hypothesis can be rewritten as $H' : N\mu + \sum n_i \alpha_i = 0$. The hypothesis is now in estimable form. Also, $K' = \lambda' = [N \ n_1 \ \dots \ n_a]$ and $m = 0$.

A testable hypothesis then is one that is made up of estimable functions. Hypotheses made up of non-estimable functions cannot be tested (for proof see in Searle 5.5d). For $K'b$ estimable and K' having full row rank ($=s$) then the test for testable hypotheses can be expressed as

$$F(H) = (K'b^0 - m)' (K'GK)^{-1} (K'b^0 - m) / \hat{s}^2 \quad (3.21)$$

$$= Q / \hat{s}^2 \text{ with } s \text{ and } N - r \text{ d.f.}$$

For the example above Q in the expression can be shown to be equal to $N\bar{y}^2$ or SSM ; also $s = r(K') = r(\lambda') = 1$ and so $F(H) = F(M)$. Thus we have shown it is possible to demonstrate that the test based on $F(M)$ in the one way classification is equivalent to testing $H: E(\bar{y}) = 0$. For the one way layout consider the case where $i = 3$, where we wish to test $H': \alpha_1 = \alpha_2 = \alpha_3$ and consequently $H'': \alpha_1 - \alpha_2 = \alpha_1 - \alpha_3 = 0$.

Expressing this hypothesis in general form we have

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (3.22)$$

where $s = r(K') = 2$,

with the expression of the form $K'b = m$, $K'b$ estimable.

It is also possible to show that $F(H)$ is equivalent to testing $F(R_m)$. Hence, the test based on $F(R_m)$ is equivalent to testing all α 's equal. At 'first thought' this might appear contradictory. If individual effects $\{\alpha_i\}$ are not themselves estimable than why can we test a hypothesis of the form all α 's equal? The reason for this apparent contradiction is that the model for all α 's equal can be expressed as

$$y_{ij} = \mu + \alpha_0 + \epsilon_{ij} = \mu' + \epsilon_{ij} \quad (3.23)$$

and the sum of squares for this model is the same as fitting $y_{ij} = \mu + \epsilon_{ij}$ (hypothesis written $H: K'b = 0$).

Sometimes a linear model may include restrictions on the elements of a parameter vector. Restrictions are considered as an integral part of the model, for example the situation $\sum \alpha_i = 0$ (or sum effects zero) is taken not as a hypothesis but as a fact without question. Such restrictions are quite different from the "usual constraints" imposed solely for obtaining a solution for the normal equations; and need have no bearing on the model whatever. Constraints (sometimes referred to as "restrictions" just to further add to any confusion, Federer (1955, p. 159)) can be perfectly permissible so long as the implications of their use are understood by the user.

As Searle cogently clarifies "these constraints cannot be 'any' constraints,... in situations of unbalanced data those of the form $\sum \alpha_i = 0$ are generally not the simplest... such constraints are not necessary for solving normal equations; they are only sufficient... they can be used whether or not a similar relationship holds in the model..."

Up to now the models discussed have not included any restrictions, the discussion has been solely in terms of the unrestricted model. It seems too easy to say that the choice of which model to use, the unrestricted model or the restricted model, depends on the nature of one's data. Clearly unquestioning acceptance of methods of "usual constraints" will fail to illuminate the impact of such assumptions on any interpretation of one's data. However, one would like to be in a position to argue for the use of a set of constraints or inclusion of specific restrictions on the basis of their intuitive appeal as well as their algebraic logic. Searle presents in the table below how different restrictions can lead to the same parameter being estimable in even though that parameter may not be estimable under the unrestricted model.

Table 3.2 : Estimators of μ and F-statistics for testing $H: \mu = 0$, in three different restricted models.

Restriction on model	Estimable function in unrestricted model which reduces to μ in restricted model	b.l.u.e. of μ in restricted model (= b.l.u.e. of function in preceding column in unrestricted model)	F-statistic for testing $H: \mu = 0$
$\sum n_i \alpha_i = 0$	$\mu + \sum n_i \alpha_i / n$.	$\bar{y}_{..}$	$F(M) = n \bar{y}_{..}^2 / \hat{\sigma}^2$
$\sum \alpha_i = 0$	$\mu + \sum \alpha_i / a$	$\sum \bar{y}_i / a$	$(\sum \bar{y}_i)^2 / (\hat{\sigma}^2 \sum n_i^{-1})$
$\sum w_i \alpha_i = 0$	$\mu + \sum w_i \alpha_i / w$.	$\sum w_i \bar{y}_i / w$.	$(\sum w_i \bar{y}_i)^2 / (\hat{\sigma}^2 \sum w_i^2 n_i^{-1})$

(Taken from Searle, 1971)

Typically for unbalanced data we find $\sum n_i \alpha_i = 0$ is used. In our context interviewer effects are weighted in proportion to their completed workload sizes and these weighted effects are assumed to cancel out across the pool of interviewers from which they have been selected. Why this should be more appealing than $\sum \alpha_i = 0$ can presumably only be justified in terms of the good sense of ensuring that interviewers who complete smaller workloads have an impact on the observations in relation to the size of that effort (and vice versa). Apart from that, something like $\sum w_i \alpha_i = 0$ might be more meaningful. Perhaps, in a design where three interviewers who had been selected from three different agencies prior to experimental allocation w_i could be chosen to reflect the proportion of interviewers to be found in the whole population of interviewers. Thus although $\sum n_i \alpha_i = 0$ provides an easy solution for \mathbf{b}^0 ($\mu^0 = \bar{y}_{..}$ and $\alpha_i^0 = \bar{y}_{i..} - \bar{y}_{..}$) the same restriction applied to the parameters of the model may not always be appropriate. Also note, under $\sum \alpha_i = 0$ the estimate for an effect becomes $\alpha_i^0 = \bar{y}_{i..} - \sum \bar{y}_{i..} / a$, i.e. different restrictions lead us to the same parameter being estimated even though it was not estimable in the unrestricted model, but the b.l.u.e.'s are not the same. Such discrepancies may or may not have later consequences for the investigator. Either restriction could be rationalised in terms of defining α 's in terms of deviation from an average, but the point here is, of course, which average to choose? (the overall mean or the mean of the interviewer means?) Obviously in the balanced case the restrictions $\sum n_i \alpha_i = 0$ and $\sum \alpha_i = 0$ are equivalent and we are protected from any anxiety.

Appropriate models for designs more complex than that for the one way layout leads us to consider the adequacy of different models for the same set of data. Indeed, even for the one way layout SSR_m is the difference between the reduction in sums of squares for fitting two models, one containing μ and an α -factor, the other containing just μ . Using Searle's $R()$ notation as mnemonic for "reduction in sum of squares" (rather than residual) $R(\mu, \alpha) - R(\mu)$ is the additional reduction due to fitting μ and α , over and above fitting just μ ; or the reduction due to fitting " α after μ ", written $R(\alpha / \mu)$. To summarize,

$$\begin{aligned}
 SSM &= R(\mu) \\
 SSR &= R(\mu, \alpha) \\
 SSR_m &= R(\alpha / \mu) \\
 \text{and} \quad SSE &= SST - R(\mu, \alpha).
 \end{aligned} \tag{3.24}$$

These mnemonics will be used below. First consider a two level nested survey design with interviewers within areas a suitable model would be

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk} \tag{3.25}$$

where μ is the mean response; now α_i represents an area effect; β_{ij} is the interviewer within area effect; e_{ijk} the residual error and y_{ijk} the response of the k -th interviewee to the j -th interviewer in the i -th area. We describe the interviewer factor as nested within the area factor. In general there would be 'a' levels of area and 'b' levels of interviewer (there is no absolute necessity to always imagine pairs of interviewers within areas); in this way

$$\begin{aligned}
 n_{i.} &= \sum_{j=1}^{b_i} n_{ij} \\
 \text{and} \\
 n_{..} &= \sum_{i=1}^a n_{i.}
 \end{aligned} \tag{3.26}$$

The sums of squares for the analysis of variance for such data are:

$$\begin{aligned}
 \text{mean} &\quad SSM = R(\mu) \\
 \text{model, after mean} &\quad SSR - SSM = R(\alpha, \beta: \alpha / \mu) \\
 \text{residual} &\quad SSE = SST - R(\mu, \alpha, \beta: \alpha)
 \end{aligned}$$

Note: $\beta: \alpha$ is a Searle convention to identify nesting.

Now suppose we fit the one way classification model to these data (ignoring any differential interviewer effect) then the reduction for fitting this model is

$$R(\mu, \alpha) = \sum_{i=1}^a y_{i..}^2 / n_{i.} \tag{3.27}$$

Subtracting this sum of squares from $R(\mu, \alpha, \beta; \alpha)$ enables us to show that the sum of squares due to fitting the model after the mean can be divided into two portions, one which summarizes the effect of fitting after the mean ($R(\alpha/\mu) = R(\mu, \alpha) - R(\mu)$), and the other which summarizes the effect of fitting β after μ and α . Applying normal theory enables us to test to what extent these factors are responsible for the variation in y_{ijk} .

Applying the general theory of estimability to this design involves many of the points already made for the one way classification. The main points of practical interest have been summarized table 3.3 below. First consider that the expected value of any observation is estimable, with b.l.u.e. $\mu^0 + \alpha_i^0 + \beta_{ij}^0 = y_{ij}$. This result and linear combinations thereof are now provided:

Table 3.3: Estimable functions in the 2-way nested classification
 $y_{ij} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$

Estimable Function	b.l.u.e.	Variance of b.l.u.e.
$\mu + \alpha_i + \beta_{ij}$	$\bar{y}_{ij.}$	σ^2/n_{ij}
$\beta_{ij} - \beta_{ij'}$, for $j \neq j'$	$\bar{y}_{ij.} - \bar{y}_{ij'}$	$\sigma^2(1/n_{ij} + 1/n_{ij'})$
$\mu + \alpha_i + \sum_{j=1}^{b_i} w_{ij}\beta_{ij}$, for $\sum_{j=1}^{b_i} w_{ij} = 1$	$\sum_{j=1}^{b_i} w_{ij}\bar{y}_{ij.}$	$\sigma^2 \left(\sum_{j=1}^{b_i} w_{ij}^2/n_{ij} \right)$
$\alpha_i - \alpha_{i'} + \sum_{j=1}^{b_i} w_{ij}\beta_{ij} - \sum_{j=1}^{b_{i'}} w_{ij}\beta_{ij'}$, for $\sum_{j=1}^{b_i} w_{ij} = 1 = \sum_{j=1}^{b_{i'}} w_{ij}$	$\sum_{j=1}^{b_i} w_{ij}\bar{y}_{ij.} - \sum_{j=1}^{b_{i'}} w_{ij}\bar{y}_{ij'}$	$\sigma^2 \left(\sum_{j=1}^{b_i} w_{ij}^2/n_{ij} + \sum_{j=1}^{b_{i'}} w_{ij}^2/n_{ij} \right)$

(Taken from Searle, 1971)

Note that μ is not estimable; neither is $\mu + \alpha_i$. The estimable functions of this table form the basis of testable hypotheses. The hypothesis of special interest is, of course, $H: \beta_{i1} = \beta_{i2} = \dots = \beta_{ib_i}$ for all i . By writing it in the form $H: K'b = 0$ it can be shown that the resulting F-statistic is equivalent to $F(\beta: \alpha/\mu, \alpha)$. It can also be used to test the equality of the β 's within each area. Regarding the restrictions that might apply,

$\sum_{j=1}^{b_i} w_{ij} \beta_{ij} = 0$ with $\sum_{j=1}^{b_i} w_{ij} = 1$ then $\mu + \alpha_i$ and $\alpha_i - \alpha_i'$ are estimable, and hypotheses about them are attractive. They ensure that the hypothesis all α 's equal is testable and independent of $F(\beta: \alpha / \mu, \alpha)$, that tests H : all β 's equal within each α -level. Another form of $w_{ij} = 1$ for all i , e.g. $w_{ij} = 1 / b_i$, still enables all area effects equal to be tested but the attractive qualities of the F-statistic disappear.

With balanced data ($n_{ij} = n$ for all i and j , and $b_i = b$ for all i) if familiar restrictions, $\sum \alpha_i = \sum \beta_{ij} = 0$, the effect is to make μ , α_i and β_{ij} individually estimable with b.l.u.e.'s $\hat{\mu} = \bar{y} \dots$, $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y} \dots$ and $\hat{\beta}_{ij} = \bar{y}_{ij..} - \bar{y}_{i..}$.

Again as in the one way case plausible effects in terms of being deviations from an average.

Extending this review to consider higher way layouts now consider a factorial design with two factors, for example interviewer and firm where interviewees might be employees within firms; the appropriate model would be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (3.28)$$

where μ is the mean response, β_j the firm effect, α_i the interviewer effect and γ_{ij} representing a firm interviewer interaction, suggesting response patterns between interviewers may not be consistent within different firms, ε_{ijk} represents residual error and y_{ijk} the observation for the k -th individual in the j -th firm for the i -th interviewer. Note that every level of one factor occurs with every level of another, unlike the nested design previously considered. In general, there are a levels of the factor with $\alpha_i = 1 \dots a$, and b levels of the β factor with $j = 1 \dots b$. With unbalanced data, when some cells have no observations (e.g. an interviewer may have fallen sick and was unable to visit one of his/her firms); there are only as many

γ -levels as there are non-empty cells. n_{ij} is the number of observations in the i,j cell in which $n_{ij} > 0$; there are s such cells.

Thus,

$$n_{..} = \sum_{i=1}^a n_{i.} = \sum_{j=1}^b n_{.j} = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \quad (3.29)$$

Once we encounter this situation we soon discover that the tidiness of the analysis of variance for a two way classification with balanced data fails to carry over to the unbalanced case with empty cells. Indeed the balanced situation provides no explanation as to why there are two analyses of variance for unbalanced data, how the manner of interpreting effects changes and why the calculations are quite different. Searle disentangles the puzzle.

Firstly consider the analysis of variance tables summarized in table 3.4 taken from Searle (p. 298). For a complete exposition of the algebraic details the reader is referred to section 7.2 (d), p.292, in particular Searle's equations (63) and (69).

Table 3.4: Equivalent expressions for sums of squares in the analysis of variance of the two way classification with interaction

Sum of Squares	Degrees of Freedom ¹	Method	
		Absorbing α 's (Use when more α 's than β 's) See (63) for $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$	Absorbing β 's (Use when more β 's than α 's) See (69) for $\mathbf{u}'\mathbf{T}^{-1}\mathbf{u}$
<i>Fitting α before β (Table 7.7b)</i>			
$R(\mu)$	1	$n_{..}\bar{y}_{...}^2$	$n_{..}\bar{y}_{...}^2$
$R(\alpha \mu)$	$a - 1$	$\sum_i n_{i.}\bar{y}_{i..}^2 - n_{..}\bar{y}_{...}^2$	$\sum_i n_{i.}\bar{y}_{i..}^2 - n_{..}\bar{y}_{...}^2$
$R(\beta \mu, \alpha)$	$b - 1$	$\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$	$\sum_j n_{.j}\bar{y}_{.j.}^2 + \mathbf{u}'\mathbf{T}^{-1}\mathbf{u} - \sum_i n_{i.}\bar{y}_{i..}^2$
$R(\gamma \mu, \alpha, \beta)$	$s - a - b + 1$	$\sum_i \sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{i.}\bar{y}_{i..}^2 - \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$	$\sum_i \sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_j n_{.j}\bar{y}_{.j.}^2 - \mathbf{u}'\mathbf{T}^{-1}\mathbf{u}$
SSE	$N - s$	$\sum_i \sum_j \sum_k y_{ijk}^2 - \sum_i \sum_j n_{ij}\bar{y}_{ij.}^2$	$\sum_i \sum_j \sum_k y_{ijk}^2 - \sum_i \sum_j n_{ij}\bar{y}_{ij.}^2$
SST	N	$\sum_i \sum_j \sum_k y_{ijk}^2$	$\sum_i \sum_j \sum_k y_{ijk}^2$
<i>Fitting β before α (Table 7.7c)</i>			
$R(\mu)$	1	$n_{..}\bar{y}_{...}^2$	$n_{..}\bar{y}_{...}^2$
$R(\beta \mu)$	$b - 1$	$\sum_j n_{.j}\bar{y}_{.j.}^2 - n_{..}\bar{y}_{...}^2$	$\sum_j n_{.j}\bar{y}_{.j.}^2 - n_{..}\bar{y}_{...}^2$
$R(\alpha \mu, \beta)$	$a - 1$	$\sum_i n_{i.}\bar{y}_{i..}^2 + \mathbf{r}'\mathbf{C}^{-1}\mathbf{r} - \sum_j n_{.j}\bar{y}_{.j.}^2$	$\mathbf{u}'\mathbf{T}^{-1}\mathbf{u}$
$R(\gamma \mu, \alpha, \beta)$	$s - a - b + 1$	$\sum_i \sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_i n_{i.}\bar{y}_{i..}^2 - \mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$	$\sum_i \sum_j n_{ij}\bar{y}_{ij.}^2 - \sum_j n_{.j}\bar{y}_{.j.}^2 - \mathbf{u}'\mathbf{T}^{-1}\mathbf{u}$
SSE	$N - s$	$\sum_i \sum_j \sum_k y_{ijk}^2 - \sum_i \sum_j n_{ij}\bar{y}_{ij.}^2$	$\sum_i \sum_j \sum_k y_{ijk}^2 - \sum_j \sum_i n_{ij}\bar{y}_{ij.}^2$
SST	N	$\sum_i \sum_j \sum_k y_{ijk}^2$	$\sum_i \sum_j \sum_k y_{ijk}^2$

¹ s = number of filled cells.

The numerical analysis involves a procedure described as "absorbing", a technique for solving the normal equations for the model either in terms of "absorbing" the β -equations and solving for (b-1) β 's or through "absorbing" the b -equations and solving for (a-1) α 's to obtain $R(\mu, \alpha, \beta)$. Details are given in Searle, section 7.1d (p 266). The computational advantages of the procedure are convincing. The two alternatives provide identical numerical results but are represented differently in terms of their symbolic identity. In the preceding designs considered the reduction sum of squares due to fitting the model simplify. Now, there is no neat solution. The possibilities in terms of interpretation of the F-statistics from the two way layout with interaction are conveniently aided by reproducing table 7.4 from Searle below.

Table 3.5: Suggested conclusions according to significance (Sig) and non-significance (NS) of F-statistics in fitting a model with two main effects (α 's and β 's).

Fitting β and then α after β					
Fit α and then β after α	$F(\beta \setminus \mu)$: $F(\alpha \setminus \mu, \beta)$:	Sig Sig	NS Sig	Sig NS	NS NS
<i>Effects to be included in model</i>					
$F(\alpha \setminus \mu)$: $F(\beta \setminus \mu, \alpha)$:	Sig Sig	α and β	α and β	β	Impossible
$F(\alpha \setminus \mu)$: $F(\beta \setminus \mu, \alpha)$:	NS Sig	α and β	α and β	β	α and β
$F(\alpha \setminus \mu)$: $F(\beta \setminus \mu, \alpha)$:	Sig NS	α	α	α and β	α
$F(\alpha \setminus \mu)$: $F(\beta \setminus \mu, \alpha)$:	NS NS	Impossible	α and β	β	neither α or β

(Taken from Searle, 1971)

First and foremost F-statistics, $F(\alpha/\mu)$, $F(\beta/\mu, \alpha)$, $F(\beta, \mu)$ and $F(\alpha/\mu, \beta)$, should only be considered if $F(R_m) = F(\alpha, \beta, \mu)$ is significant. Otherwise simultaneous fitting of both α and β have little explanatory value for the variation in y_{ijk} . Of course, it still does not mean that both α and β are needed in the model. Examination of the F-statistics above casts light on this aspect of the model. In addition, we have the statistic $F(\gamma / \mu, \alpha, \beta)$ which provides a test of the effectiveness of fitting the current model against the no interaction model. The hypothesis tested by the F-statistic is difficult to unravel; essentially there are $s - a - b + 1$ degrees of freedom of $R(\gamma / \mu, \alpha, \beta)$ which tests hypotheses relating to column vectors of linearly independent functions of

$$\theta_{ij}, i'j' = \gamma_{ij} - \gamma_{i'j'} - \gamma_{ij'} + \gamma_{i'j} \quad (3.30)$$

where such functions are either estimable or estimable sums or differences of θ 's. It is difficult to see how these expressions might have any applied meaning in the presence of interaction unless one had sound a priori reasons for expecting differences between different pairs of interviewers working in different firms.

In general, the form of the basic estimable function for the two way classification with interaction is $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ with b.l.u.e. $\hat{\mu}_{ij} = \mu^0 + \alpha^0_i + \beta^0_j + \gamma^0_{ij} = \bar{y}_{ij}$.

μ_{ij} is only estimable if the corresponding (i, j) cell contains observations. Any linear function of the μ_{ij} is estimable, but because of the presence of γ_{ij} in μ_{ij} , differences between interviewers (α_i) or firms (β_j) are not estimable.

Therefore $\mu_{11} - \mu_{21} = \alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21}$ is estimable but $\alpha_1 - \alpha_2$ is not. In general,

$$\alpha_1 - \alpha_{i'} + \sum_{j=1}^b k_{ij} (\beta_j + \gamma_{ij}) - \sum_{j=1}^b k_{i'j} (\beta_j + \gamma_{ij}) \quad (3.31)$$

for $i \neq i'$ is estimable so long as $\sum_{j=1}^b k_{ij} = \sum_{j=1}^b k_{i'j}$ with $k_{ij} = 0$ when $n_{ij} = 0$ and $k_{i'j} = 0$ when $n_{i'j} = 0$.

Similar results hold for the β 's. Note the γ 's are always involved thereby handicapping the opportunity of constructing an estimable function solely in terms of either the α 's or the β 's.

The remainder of the hypotheses tested by the F-statistics are summarized below:

(i) F(M) tests $H: \sum_{i=1}^a \sum_{j=1}^b n_{ij} \mu_{ij} = 0 \text{ for } n_{ij} \neq 0$
(equivalent to $H: E(\bar{y}) = 0$)

(ii) $F(\alpha / \mu)$ and $F(\beta / \mu)$ test $H: \sum_{i=1}^a \sum_{j=1}^b n_{ij} \mu_{ij} \text{ equal for all } i$

(iii) $F(\alpha / \mu, \beta)$ and $F(\beta / \mu, \alpha)$ test
 $H: \theta_i = 0 \text{ for all } i$
and $H: \psi_j = 0 \text{ for all } j$

Exact forms of θ_i and ψ_j are given in Searle, equations (86) and (87) p.304. They are complex expressions involving weighted sums and differences of α 's and γ 's. They are not immediately pleasing as far as interpretability is concerned. The form of (3.31) suggests $\alpha_i - \alpha'_i$ will be estimable if the model includes the restrictions $\sum_{j=1}^b k_{ij} (\beta_j + \gamma_{ij}) = 0 \text{ for all } i \text{ for } n_{ij} \neq 0$.

In particular, if $k_{ij} = n_{ij} / n_i$, this becomes

$\sum_{j=1}^b n_{ij} (\beta_j + \gamma_{ij}) = 0 \text{ for all } i \text{ for } n_{ij} \neq 0$ and the corresponding b.l.u.e. is $\bar{y}_{i..} - \bar{y}_{i'..}$.

The problem with having such restrictions as part of the model is that they are data dependent. Both are functions of the n_{ij} and which of the cells are non-zero. For data having all cells filled the situation is more optimistic, the estimable function becomes a special case of (3.31)

with $\alpha_i - \alpha'_i + (\sum_{j=1}^b \gamma_{ij} - \sum_{j=1}^b \gamma'_{ij}) / b$ such that $k_{ij} = k'_{ij} = 1 / b$.

Also the joint hypothesis

$$H: \alpha_i + \sum_{j=1}^b \gamma_{ij} / b \text{ all equal for } i = 1 \dots a \quad (3.32)$$

can also be tested so that if the model includes the restriction $\sum_{j=1}^a \gamma_{ij} = 0 \text{ for all } i = 1 \dots a$, then the hypothesis reduces to (refer to section 7.2 (h), Searle) testing equality of all α 's. Similar results hold for the β 's.

In conclusion, for the most general case of the two way classification with unbalanced data and empty cells we have unattractive hypotheses that are dependent on the structure of the available data.

For instance, the hypotheses often involve the n_{ij} 's not only in terms of the weight attached to an effect, but also in relation to whether some of the effects enter the hypotheses at all. As Searle suggests "usually an experimenter wishes to test hypotheses that arise from the context of his work and not on hypotheses that depend on the pattern of the n_{ij} 's in the data". In general this cannot be realised; though there may be some hope if we are dealing with proportionate subclass numbers, as may be the situation in a carefully designed factorial. Then using the two way classification with no interaction we have $p_j = n_{ij} / n_{i.}$, which would be equivalent saying every interviewer completes the same number of interviews in each firm. If so, the corresponding F-statistic becomes equivalent to testing $H: \alpha_i$ equal for all i .

Urquhart et al (1970) in their revisit of estimation problems associated with linear models suggest that as linear models have become more widely utilized they have simultaneously become less well specified. Specifically they have become "over parameterized" - they contain more parameters than is necessary to describe the experimental context. In turn experiments do not support their estimation, for example in the two way classification we had s linearly independent means but $(1 + a + b + s)$ parameters. The idea of estimability was introduced to circumvent this problem.

Their proposal considers an alternate linear model which is more closely identified with the experimental context, namely the μ_{ij} model. Basically it is assumed that the experimenter has sampled s different populations for the purpose of studying relationships among the means of these populations. These s populations generate s observed sample means; each of which is an estimator of the population mean from which the original observations are deemed to be a sample. These population means are estimated without definitional ambiguity. Consider the two way classification with interaction, the model is written simply as

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad (3.33)$$

where the ε_{ijk} have the same distributional properties as before. Then μ_{ij} has b.l.u.e. \bar{y}_{ij} , with $V(\hat{\mu}_{ij}) = \sigma^2 / n_{ij}$.

Elegant simplicity arises since there is no confusion as to what functions are estimable, what their b.l.u.e.'s are and what hypotheses can be tested, the models are always of full rank. It is up to the investigator to specify functions and hypotheses of interest in terms of the μ_{ij} . Estimating linear functions of μ , expressed as $t'\mu$ or $T'\mu$ present no problem.

The outstanding problem posed, of course, is the selection of an interesting set of linear functions of vector means. This is no easy task, though it may come as some comfort to consider the view expressed by Urquhart et al (1970) that the question of estimability of certain linear functions of parameters in linear models is a result of statisticians failure to be precise in assisting experimenters in choosing T 's for their problems. Clearly the investigator will have a much easier time interpreting t' when s/he picks t' so that t' has a fairly obvious relation to the concerns of the investigation rather than abdicating responsibility for selecting t' to some arithmetic process s/he may not understand.

Formally, the basic formulation of the μ_{ij} -model assumes s sampled univariate populations, each population having a mean (here μ_{ij}) where certain restrictions may or may not be known about the means

$$\begin{matrix} y = & W & \mu & + \epsilon \\ & nx1 & nxm & mx1 \end{matrix} \quad (3.34)$$

(with $E(\epsilon) = 0$, $\text{cov } (\epsilon) = V$ and $P'\mu = c$ as a restriction) and,

$$W = \begin{matrix} & n_{\alpha_1} & & & \\ j_1 & & & & \\ & & n_{\alpha_2} & 0 & \\ & & j_1 & & \\ & & & & n_{\alpha_m} \\ & 0 & & j_1 & \end{matrix}$$

where $j_1 n_{\alpha_k}$ is a vector of 1's determined by the number of observations in each cell of the classification.

In addition some authors (Speed, 1969) assume each population has a common variance such that $\text{cov}(\epsilon) = \sigma^2 I_n$. In either the unrestricted or restricted case the function $t'\mu$ has an unbiased

$$\text{estimate, namely } E(t'\hat{\mu}_u) = t'\mu = E(t'\hat{\mu}_r) \quad (3.35)$$

However, the two estimates may not have the same variance (Urquhart et al, 1970). This turns out to be a benefit for μ_r in that it may turn out to have less variance than the corresponding elements of μ_u . The benefit arises due to having additional information about elements of μ . Urquhart considers the following example for a two way classification (2x3), where we will imagine two interviewers have been assigned to three interviewing locations, each interviewer interviewing in every location (e.g. a firm) such that

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ -1 & -1 & 2 & -1 & -1 & 2 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 1 & -2 & -1 & -1 & 2 \end{bmatrix} * \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} \quad (3.35)$$

Now $\hat{\mu}_u = (D^{-1}W'y)' = \bar{y}'$, where D is a diagonal matrix with elements n_{ij} , and consequently $T'\hat{\mu}_u = \hat{\theta}'_u$. Now suppose that the researcher knows $P'\mu = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$ then it is possible to deduce $\hat{\mu}_r$ and $(T'\hat{\mu}_r)'$. The main issue relates to the choice of T and the knowledge leading to such a restriction. At 'face value' such restrictions may make little sense. Searle (1986) suggested that by including such a restriction the researchers were also assuming that $\mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} = 0$, which would amount to a 'zero' interaction assumption when taken together. But here, we come full circle having left traditional analysis of variance due to 'over parametrization' we return to its' conventions to guide our choice of sensible functions of μ . For the example in (3.35)

θ_1 is associated with the overall mean

θ_2 is associated with the main effect of interviewer

θ_3, θ_4 are associated with the main effect of location, and

θ_5, θ_6 are associated with interviewer-location interaction.

The usual procedure for testing for zero interaction involves (equivalently) testing both θ_5 and θ_6 simultaneously zero. Decomposing the interaction into components provides more information. Note that the conventional 2-way equivalent hypotheses defined average interaction effects to be zero.

Speed (1969) presents a thorough account of this whole topic. Indeed he has developed a theorem whereby it is possible to find the appropriate μ_{ij} -model given a conventional analysis of variance formulation. The reader is referred there for further elucidation. What remains is the attractive merit of the simplicity of the approach where responsibility for expressing meaningful restrictions and functions of the μ lie with the researcher. The approach also has appeal in the sampling context, if, instead of conceptualizing the six populations above as populations in their own right they were described as strata of a larger population then this might lead to the choice of elements of T' that related to the sizes of various strata. Searle touches on this point in section 8.1f (p.339) when discussing the analysis of large scale survey-type data, the cell means are referred to as sub-most cells, cells defined by one level of each of the stratification factors. Just what hypotheses get tested are the prerogative of the person whose data they are. Nevertheless, the author agrees with Searle (1987) that perhaps the time is ripe for learning about linear models, not in terms of difficulties incurred with over parametrized models but through using easier and (potentially) more informative cell means models for unbalanced data. Searle's 1987 publication deals largely with the fixed effects model and is an appropriate celebration of the model properties ideas expressed in this section. We now go on to consider the inclusion of random effect.

3.5:

What happens to the linear model if some or all of the effects are random?

Under the fixed effect assumption interviewer effects are regarded as fixed in relation to each other. From a sampling view data is envisaged as one possible data set involving these same interviewers that could be derived in repetitions of the survey design. The residual error terms are considered as a random sample from a population of error terms, such that $e \sim (0, \sigma_e^2 I)$. It is the randomness of the e 's that provides the means for making inferences about functions of the interviewer (and other fixed) effects α 's and e 's.

In the random effects model the concept of error remains the same but the effects are considered to be a random sample from a population of effects, so for a one way layout $\alpha \sim (0, \sigma_\alpha^2 I)$. Further, the sampling of α 's is assumed to be independent of the e 's, the variance of an observation can be expressed as $\sigma_y^2 = \sigma_\alpha^2 + \sigma_e^2$. Following Kish (1962) when using ρ as a summary measure of any interviewer effect necessarily leads us to consider the estimation of variance components and inferences about them.

Extending the approach to embrace higher way layouts as in the previous sections (and 2.2) raises the possibility that not all of the effects in a model need to be considered as either fixed or random. Models may be defined in terms of a mixture of effects or as mixed effects models; for example, in the two way classification for interviewer by firm/location described in the previous section firm could have remained as a fixed effect and interviewer considered as random.

It is important to emphasise that the assumption of random effects does not carry with it the assumption of normality; if our interest is in the distributional properties of estimators then the normality of random effects is typically assumed.

In the case of balanced data the procedure for estimating components relies on one method, that is to derive expected mean squares as though the underlying model were fixed and equate them to calculated (observed) values. This leads to

linear equations for variance components. The mean squares in analysis of variance are quadratic forms of the observations, an important result in this regard when $x \sim N(\mu, V)$

then $E(x' Ax) = \text{trace } (AV) + \mu' A \mu$ (3.36)
 (true also when x non-normal)

for any matrix A (see Searle, 1971, chapter 2).

In fixed effects models $\text{var}(y)$ has been of the form $\sigma_e^2 I$. This is not the case for random effects because the covariance structure of random effects determines the variance-covariance matrix of observations,

essentially, $\text{var}(y) = \sum_{i=1}^a (\sigma_e^2 I_i + \sigma_\alpha^2 J_i)$ (3.37)

where J is a square matrix every element unity: Where I and J have order equal to the number of observations in each group.

the corresponding form for unbalanced data is

$$\sum_{i=1}^a (\sigma_e^2 I_i + \sigma_\alpha^2 J_i) \quad (3.38)$$

where I_i and J_i are of order n_i

The procedure of equating mean squares to expected values is a special case of a general procedure of equating quadratic forms to their expected values as used in a variety of ways with unbalanced data. For balanced data the "obvious" quadratic forms are analysis of variance mean squares, the resulting estimators can be easily obtained utilizing a set of rules, due principally to Henderson (1959, 1969) and described in Searle (section 9.6 p. 389, 1971). These rules apply to crossed and nested designs; partially balanced and Latin Square designs are excluded.

One important observation relating to the application of these rules relates to the mixed effects model. For example, consider the fixed effect as firm j denoted by β and the random effect is interviewer i by α , where γ denotes the interaction term.

For the balanced model

$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$, where $i = 1..a$, $j=1..b$ and $k=1 ...n$, then with no restrictions on the interaction effects

$$E(\text{MSA}) = \frac{bn}{a-1} \sum_{i=1}^a (\alpha_i - \bar{\alpha})^2 + n\sigma^2_{\gamma} + \sigma^2_e \quad (3.39)$$

alternatively, with the restriction $\sum_{i=1}^a \gamma_{ij} = \gamma_{..j} = 0$ for all j this leads to a slightly different anova table, where

$$E(\text{MSA}) = \frac{bn}{a-1} \sum_{i=1}^a (\alpha_i'' - \bar{\alpha}'')^2 + n(a / a-1) \sigma^2_{\gamma} + \sigma^2_e \quad (3.40)$$

where " notation denotes same structure but different restrictions on γ 's.

Searle points out that this dual approach to the mixed model is evident in many places (e.g. Anderson and Bancroft, 1952, p. 339) but he, himself, prefers the former approach as the results are consistent with those for unbalanced data. There is a lengthy discussion on the choice of restriction in Wilk and Kempthorne (1955, 1956).

Before detailing the methods of estimation appropriate to unbalanced data it will be useful to summarize the properties of the estimators. These properties have a disconcerting tale to tell for unbalanced data:

(i) unbiasedness and minimum variance whether the model is mixed or random variance component estimators are unbiased (in fixed or mixed models the "fixed" part is not used in estimating variance components). Unbiasedness does not hold for estimation procedures for mixed models with unbalanced data. Minimum variance properties, with or without normality, apply in the unbalanced case but are limited to one way classifications (Townsend, 1968; Harville, 1969).

(ii) negative estimates: by definition a variance component is positive; unfortunately estimates can be negative; this is so not only in the simple one way balanced layout but in more sophisticated designs both for balanced and unbalanced data. Searle (1971) suggests a few courses of action: accept as

evidence that the true component is zero, retain as negative in the anova table or use as zero (consequently upsetting the unbiasedness property), use as indication of zero component and re-estimate the other components using a method of "pooling minimal mean squares with predecessors" (Thompson, 1961, 1962.)

Other alternatives are more radical, they involve questioning the appropriateness of the model, questioning the appropriateness of the method that yielded it (possibilities are to use Bayes procedures, Tiao and Box (1967), or maximum likelihood estimators, Herbach (1959) and Thompson (1962)) or collect more data (the statistician's last hope!).

Negative estimates are solely a function of the estimation procedure and data. If normality is assumed it is possible to derive the probability of obtaining a negative estimate.

Normality assumptions for the error terms and every set of random effects in the model bring the following delights:

(a) distribution of mean squares will be central chi-square and non-central chi-square for terms involving fixed effects. This result is demonstrated for the balanced case in Searle by expressing mean squares in quadratic form ($y' Ay$) and utilising theorems developed in his chapter 2 (1971). The steps have important generalizations for unbalanced data.

(b) distribution of estimators; even in the balanced one way classification the exact form of the distribution of (akin to the interviewer variance component) cannot be derived. For a interviewers and workload size of n each we have:

$$\sigma_{\alpha}^2 \sim \frac{n\sigma_{\alpha}^2 + \sigma_e^2}{a(n-1)} \chi^2(a-1) - \frac{\sigma_e^2}{an(n-1)} \chi^2(an-a) \quad (3.41)$$

Unknown values σ_e^2 and σ_{α}^2 occur in the coefficients and the second term is negative. Were the coefficients known techniques exist for the solution (Robinson, 1965; Wang, 1967). This result holds generally. The distribution for σ_e^2 can be defined exactly (see Searle, p. 410).

(c) tests of hypotheses: In the random effects model all ratios of mean squares have central F-distribution; in mixed models the same is true of ratios of mean squares whose expected values contain no fixed effects.

However Searle (1971) shows that the table of "expected values" will not always suggest the "obvious" denominator for testing a hypothesis, he does this by utilizing Satterthwaite's procedure (1946) for testing $H: \sigma^2_c$ where σ^2_c is any component in the model.

(d) confidence intervals; being unable to derive exact distributions does not necessarily preclude the possibility of deriving approximate intervals (Graybill, p. 369, 1961), in some cases the intervals are exact, the most obvious being for σ^2_e , based on the chi-square distribution. Other exact intervals are readily available for the one way balanced layout (see Searle, table 9.14, p. 414). A result of particular interest here is the one for the ratio of variance components,

$\sigma^2_a / \sigma^2_a + \sigma^2_e$, namely ρ . The interval estimation procedure is based on work by Scheffe (1959), Graybill (1961), Williams (1962) and Broemling (1969). The suggestion was not utilized in subsequent evaluations due to the computational ease of 'Jackknife' repeated replication (see chapter 7).

(e) probability of negative estimates. Leone et al (1968), p. 415, provide a procedure for determining the probability of a variance component being negative (generally where $\sigma^2_e = (m_1 - m_2) / k$; m_1, m_2 are mean squares). For the one way balanced layout this provides

$$P_r(\sigma_a \text{ negative}) = P_r\{F_{a-1, a(n-1)} < 1 / 1 + n\rho\} \quad (3.42)$$

$$\text{where } \rho = \sigma^2_a / \sigma^2_e$$

Note the procedure requires giving values to the components σ^2_a and σ^2_e . Searle (1971) suggests using a series of arbitrary values to obtain such an indication.

(f) sampling variances of estimators. Although the distribution functions of estimators cannot generally be derived, sampling variances of variance component indicators can be. The problem is that the variances are functions of the unknown components. With balanced data the mean squares are independent with known distributions and can be easily derived (see Searle, p. 416); for unbalanced data Searle uses the fact

that the estimators are linear functions of mean squares which are themselves quadratic forms of the observations, and applies his theorem 1 of chapter 2 (see Searle Chapters 10 and 11). The resultant expressions are complex.

(g) maximum likelihood estimation. In many cases under normality assumptions estimating parameters of a fixed effects model by the method of maximum likelihood leads to the same estimators as do the methods of least squares and best linear unbiased estimation. The same is not true for variance component estimation; the possibility of negative estimates for analysis of variance estimators shows they cannot be maximum likelihood estimators. The parameter space over which the likelihood is maximised has to be non-negative so far as variance components is concerned. In the early 1970's the problem of deriving maximum likelihood estimates for variance components was not as straightforward as for fixed effects situations. Indeed explicit estimators for estimators in the unbalanced case could not be found, see Herbach (1959) and Thompson (1962) (also Searle table 9.15, p.419). We shall return to maximum likelihood estimation in section 3.6 .

Estimating variance components from unbalanced data in contrast to balanced data relies on several possible procedures. We have seen in section 3.4, and tables 3.4, that with unbalanced data for the two way layout with interaction there are two analyses of variance, one for fitting α before β and the other for fitting β before α . Thus there is no uniquely "obvious" set of sums of squares or quadratic forms that can be used for variance estimation. We lack criteria for choosing between them. As Searle suggests "there is instead a variety of quadratic forms that can be used..."

Using the general result for the quadratic form $y'Ay$ Searle shows how this applies successively to a fixed, mixed and a random effects model. The methods have the advantage that the estimates have no distributional requirements. Searle's presentation is largely in terms of a two way classification with interaction. Such a model is considered as the simplest one to display most of the difficulties to be encountered with unbalanced data.

There are three principal methods of estimation:

(i) analysis of variance method or Henderson's method 1: essentially this procedure mimicks the analysis of variance method for balanced data, consisting of equating mean squares to their expected values.

The expressions used have been solely established by analogy with the balanced case. However sums of squares are not always positive in the unbalanced case - so they are not strictly sums of squares - hence the more humorous description of the procedure as the "analogous analysis of variance method" (Searle!). Expectations for estimating the variance components can be obtained from the theorem mentioned above or by direct substitution of the terms of the model into the expressions for the expected mean squares (the "brute force" method). Whatever problems emerge squares contain functions of the fixed effects that cannot be eliminated.

There are two ways of overcoming the problem either ignore the fixed effects or assume them to be random, but both are regarded as unsatisfactory as they result in biased variance component estimates.

(ii) Henderson's method 2: this procedure was designed to be suitable for mixed models. The method (Henderson, 1953) first uses the data to estimate fixed effects, these estimators are then used to adjust the original data and the variance component estimates obtained from the resulting adjusted data. The method leads to unbiased estimators (some controversy here, having said method was not uniquely defined in 1971, Henderson, Searle and Schafer, 1974, said it was). Additionally the sums of squares terms under normality assumptions do not have chi-square distributions, nor are they distributed independently of one another. The only exception is the error sum of squares. Despite this variances of estimators can be obtained under normality assumptions.

(iii) fitting constants model or Henderson's method 3: Fitting linear models described in the previous section can be referred to as the method of fitting constants, because fixed effects could be regarded as constants. This method adopts a method which uses reductions in sums of squares similar to that outlined in section 3.4 by fitting sub-models of the full model under consideration. Any computed reduction in sums of squares is equated to its expected value under the full model.

The fundamental characteristic of the method is that the parameter vector of the full model is partitioned into two components, one containing just random effects and the other fixed effects. This facilitates a major advantage for applications of the mixed model; variance component estimates that are unbiased by the presence of fixed effects. Its principal disadvantage is that it requires matrix inversion that becomes particularly difficult for models involving more than two factors. The remaining difficulty is, of course, which sums of squares to use? Searle surveys this dilemma in an elegant manner for the two way classification with interaction (p. 446 to p. 449) where we see clearly how, depending on whether you fit ' α after β ' or ' β after α ', estimators for σ^2_α and σ^2_β are the same in both instances but they are not for σ^2_β and σ^2_α . Searle suggests there is almost no answer to this problem. In the fixed effects model there might be good reason for fitting ' α after β ', or vice versa but not so in the random effects model. For four variance components in the two way classification there are five equations for estimation. Higher way layouts result in the availability of many reduction sums of squares. One way of overcoming this problem was suggested by Robson (1957); by representing all possible reductions by a vector r then

$r = A\sigma^2$ are equations we would like to solve for σ^2 , estimates for σ^2 could be obtained by least squares provided elements in r are LIN and A has full column rank ($\sigma^2 = (A'A)^{-1}A'r$).

Finally on the basis of normality assumption sampling variances of variance components can be obtained as the estimators are linear combinations of the reductions. These manifestations involve cumbersome matrix manipulation (Searle p. 451).

We have seen how data in which every subclass of the model contains observations can in fixed effects models can be analysed in terms of the means sub-most cells.

The mean squares for such analyses can also be used for estimating variance components in random or mixed effects models. This analysis of means either involves taking unweighted means as though they were observations with uniform sampling error or taking weighted means, i.e. weight terms in sums of squares in inverse proportion to the variance of the term concerned. The latter approach leads to exact F-tests (refer table 10.3 Searle, p. 452). The resulting estimators are unbiased.

A further method suggested by Koch (1967, 1968) "the symmetric sums" method estimates variance components on the basis of symmetric sums of products of the observations rather than sums of squares. The method is feasible because expected values of these products are linear functions of variance components. Means of these products produce unbiased estimators of the components. In some instances Koch also uses symmetric sums of squares of differences.

For any random (or mixed) model with balanced data the method of estimating variance components relies equating the expected values of the mean squares to their observed values for the corresponding fixed effects model. Mean squares in analysis of variance are quadratic forms in the observations. Searle refers to this procedure as the "analysis of variance method". With unbalanced data we have seen that there are many quadratic forms of observations that can be used for estimating variance components. Each of the methods reviewed simplifies to the analysis of variance method whenever data is balanced. There are no criteria for selecting criteria to choose between which quadratic forms to use. Searle suggests that these methods yield universally unbiased estimators for random effects models or the fitting constants method for mixed effects models. Though he goes on to question the value of the property of unbiasedness in this context: essentially its presence is borrowed from fixed effects estimation under notions of "repeatability". With unbalanced data from random models this idea of may be acceptable but may not result in the same pattern of unbalancedness or the same set of random effects. Alternative criteria, such as "model unbiasedness" has been suggested (Searle, 1968), yet remain cautioned (Harville, 1969).

In choosing between the various methods described in this section comparisons may be considered that focus on the sampling variances of the estimators themselves (as all methods are equivalent under balanced data conditions). However the situation soon becomes "murky", variances are themselves functions of variance components (as with balanced data) and variances are only tractable if "normality" is assumed. Comparisons have therefore largely been numerical evaluations (e.g. Bush and Anderson (1963) examined a two way classification model with interaction under several planned allocations of n). The difficulty that arises with numerical assessment is, of course, how to plan unbalancedness with enough generality to have wide applicability of findings. I leave the reader with Searle's advice: "the analysis of variance

method commends itself because it is the obvious analogue of balanced data and easy to use, though some of its terms are not sums of squares and it gives biased estimators in mixed models. The generalized form of Henderson's method 2 makes up for this deficiency but is not uniquely defined. The fitting constants method uses sums of squares that have non-central distributions in fixed effects models; it gives unbiased estimation in mixed models but it can involve more quadratics than there are components to be estimated. When all cells are filled the analysis of means has the advantage of being easy to compute, especially for unweighted means."

Each classification scheme or model considered so far has at least a general mean, a fixed effect, and an error term, a random effect, thus use of the term "mixed" could cover all modelling approaches. The importance of the fitting of constants method is its appropriateness for mixed models (Searle, p. 445). So it may suggest itself as an unifying procedure for the analysis of unbalanced data. However it is only a method for providing unbiased estimates of components of variance and gives no guidance on how to estimate fixed effects. Where variance components are known there is no problem, but typically they are not and we have the problem of wanting to simultaneously estimate both the fixed and variance components of the model. Searle suggests two courses of action:

- (i) use the fitting of constants method to estimate variance components then substitute these estimates for the true components in the generalised least squares equations for the fixed effects or
- (ii) estimate fixed effects and variance components simultaneously with a unified procedure such as maximum likelihood.

Both courses of action result in iterative procedures. Attention will be focussed on maximum likelihood procedure simply because Searle's observations began to mark a relative "burst" in its popularity as a means of overcoming the estimation difficulties associated with analysis of variance methods together with accelerated advances in computational power. However the volume and complexities of contingent algebraic forms witnessed with analysis of variance methods for unbalanced data are not jettisoned by the appearance of a unifying procedure. Searle (1971) demonstrates that even for a 1-way classification model with

$$V = \text{var}(y) = \sigma_e^2 I_N + \sigma_\alpha^2 \sum_{i=1}^a J_{n_i} \quad (3.43)$$

and a likelihood function, on the basis of normality, that no explicit solutions can be given for μ , σ_e^2 and σ_α^2 (p. 463, Searle). Indeed "explicit MLE" was despaired of by Searle. Hartley and Rao (1967) developed a general set of equations from which specific estimates are obtained by iteration, and Searle (1970) derives large sample variances for MLE variance components despite the absence of explicit estimators. For the original detail of the algebraic complexity see Hartley (1967); of course, equations reduce to simpler forms in the balanced case (mildly comforting but hardly pertinent to application). The mixed model also has simplifying features when only one factor is random. The problem of "global maxima" (or to what extent do starting values in iterative solutions affect final solutions?) or that of obtaining non-negative estimators for components remain today.

In the following section we explore more recent advances in the application of powerful "Maximum likelihood" and 'Restricted maximum likelihood' algorithms that provide simultaneous estimates of variance components and fixed effects.

3.6:

Maximum likelihood estimation

3.6.1:

The development of variance component analysis during the last two decades

It is convenient to pursue recent history in variance component analysis via an example on teaching styles presented by Aitken et al (1981). The illustration is important because it demonstrates one of the first applications of unbalanced variance component ("mixed") models in educational research by means of the EM algorithm. This motivated and allowed subsequent application to interviewer effects experiments. The authors present a two level nested (hierarchical) design, reproduced in summary form here:

Y_{pqr} denotes an achievement score for the r -th child in the q -th classroom by teaching method p ($r = 1 \dots n_q$)

x_{pqr} is a pre-test score

$q = 1 \dots Q$ (in e.g., 36), $p = 1 \dots P$ (in e.g., 3) and $N = \sum_q n_q$ (921, in e.g.)

It is assumed that teachers are randomly chosen from a population of teachers, randomly assigned to classrooms and teaching method at the beginning of a school year, where children are randomly divided into classes of roughly equal size (for further discussion regarding departures from these assumptions see original text).

The model becomes

$$Y_{pqr} = \mu + \gamma x_{pqr} + \alpha_p + T_q + E_{pqr} \quad (3.44)$$

where T_q and E_{pqr} are independently normally distributed random variables (i.n.d.r.v.) with $T_q \sim N(0, \sigma_T^2)$ and $E_{pqr} \sim (0, \sigma_E^2)$. The α_p are constants or "fixed" effects, with $\alpha_3 = 0$ (or $\sum \alpha_p = 0$), presenting mean achievement differences between methods 1 and 2, and method 3. The T_q 's are regarded as the "ability" of the q th teacher; they are treated as random variables rather than fixed constants.

The properties of such a model are discussed in Searle, (1971, chapters 9 and 10.) An analogy between teachers for interviewers, method for training, in this illustration and an "interviewer effects" study might be to substitute experience and/or agency, child for respondent, pre-test score for a previous interview response. In the discussion below, "teaching ability" might be thought of as 'interviewing style'.

Indeed, as the authors report a consequence of random teacher effects is that the achievement scores of children within the same classroom will be positively correlated:

$$\begin{aligned}
 \text{Var}(Y_{pqr}) &= \text{var}(T_q + E_{pqr}) = \sigma_T^2 + \sigma_E^2 \\
 \text{Cov}(Y_{pqr}, Y_{pqr}') &= \text{cov}(T_q + E_{pqr}, T_q + E_{pqr}') \quad (3.45) \\
 &= \text{var}(T_q) = \sigma_T^2 \\
 \text{Corr}(Y_{pqr}, Y_{pqr}') &= \rho \\
 &= \sigma_T^2 / (\sigma_T^2 + \sigma_E^2)
 \end{aligned}$$

Thus, ρ will be zero, whenever $\sigma_T^2 = 0$. For unequal "class" (read workload) sizes, they confirm that efficient estimators of the variance components, and of the fixed effects, can be obtained by maximum likelihood in the unbalanced case at the expense of considerable computation.

In their final report (ESRC, HR5710) Aitken et al., provide several approximate analysis of variance methods, each giving conflicting answers. This resulted in an application of GENSTAT for maximum likelihood estimation based on the EM algorithm (Dempster et al., 1977).

Essentially, the model described above is expressed as

$$Y/T \sim N_N(X\beta + WT, \sigma_E^2 I_N) \quad (3.46)$$

where $T \sim N_Q(0, \sigma_T^2 I_Q)$

Y is the N -vector of observations, β the vector of regression coefficients of the 'fixed' effects, of dimension r , X is the $(N \times r)$ design matrix of the fixed effects, of rank r , T is the unobserved vector of abilities of Q ($=36$) teachers and W is the $N \times Q$ design matrix for T .

The unconditional distribution of Y is multivariate normal with $E(\beta) = X\beta$, $V(Y) = \sigma^2 H$, where

$$\begin{aligned}\sigma^2 &= \sigma^2_E, H = I + \gamma WW', \\ \gamma &= \sigma^2_T / \sigma^2_E\end{aligned}\tag{3.47}$$

The maximum likelihood estimates of β , σ^2 and γ are found by differentiating the log-likelihood of Y . The likelihood equations given by Hartley and Rao (1967) are not immediately soluble, thus an iterative procedure was necessary. Hemmerle and Hartley (1973) and Thompson (1975) provide the computational details for reducing the amount of work necessary to solve these equations. The EM algorithm can be used to yield an iterative procedure, see Dempster et al., (1977).

Here 'teacher abilities' are regarded as "missing data". If these had been observed, then the maximum likelihood estimates of β , σ^2_E and σ^2_T would be

$$\beta = (X'X)^{-1} X' (Y - WT)$$

$$\begin{aligned}N\sigma^2_N &= (Y - X\beta - WT)' (Y - X\beta - WT) && \text{the 'E' step} \\ &= (Y - X\beta)' (Y - X\beta) - 2 (Y - X\beta)' WT + T' W'WT\end{aligned}$$

$$Q\sigma^2_T = T'T$$

Thus, the sufficient statistics involve the unknown T , through T , $T'T$ and $T'W'WT$, which are in the E-step, replaced by their conditional expectations given the observed data Y . These are obtained from the conditional distribution of T given Y , which is

$$T|Y \sim N_Q(\gamma W H^{-1} (Y - X\beta), \gamma \sigma^2 (I_Q - \gamma W H^{-1} W))\tag{3.49}$$

Expressions for $E(T/Y)$, $E(T'T/Y)$ and $E(T'WWT)$ follow (see p442., Aitken et al., (1981) for detail). The algorithm begins with initial estimates of β , σ_E^2 and σ_T^2 , which are then substituted in the expressions for the conditional expectations (the M-step). These conditional expectations are then resubstituted in (3.48) to give new parameter estimates until convergence occurs in 6 to 12 iterations, starting with $\gamma=1$.

Maximum likelihood (ML) estimators of variance components are biased. Patterson and Thompson (1971) developed Restricted maximum likelihood or REML, to ensure unbiased estimates of variance components. Estimation of the variance components is restricted to the error subspace orthogonal to the ML estimate of β , the "fixed" effects vector. An advantage of this procedure is that it only estimates β , once. An EM algorithm for REML is only a slight modification of the one described in (3.48) and (3.49).

Aitken et al used GENSTAT (1981) for both ML and REML estimation - slight differences were said to be observed, but both methods led to similar conclusions. REML estimates were reported for all parameters. An alternative general mixed model programme, BMDP- 3PV (1985), was not available to the authors.

Again, applying GENSTAT Anderson and Aitken (1985) developed an application to interviewer variability for binary responses (opening up methodology to non-normal applications) for an unbalanced variance component model for a two-level nested or interpenetrating design (for instance a random allocation of respondents to interviewers within geographical areas). In a wider context this example represents one of many recent applications of the methodology in fields other than agricultural or animal experimentation. Harville (1977) recognised such potential with useful survey data applications following the formulation of the EM algorithm, e.g. in addition to Aitken et al (1981), see Dempster, Rubin, Tsutakawa (1981) and Mason, Wang, Entwistle (1984) for illustrations.

Computational difficulties may still arise when the number of levels or random coefficients is large because inversion of very large matrices is required, so there are severe limitations on the size of practical problems that can be handled. Also the EM algorithm does not provide standard errors for the estimates.

Goldstein (1986) and Longford (1987) have independently constructed algorithms which do not require large matrix inversion and converge rapidly. Whilst EM methodology spurred their conception they are not actually based on the EM algorithms. Longford (1987) presents details of a Fisher scoring algorithm for the unbalanced nested random effects model using explicit formulae for the inverse and determinant of the covariance matrix, given in LaMotte (1972). Interactive software based on this algorithm was used in Aitken and Longford (1986), (for further detail on software see Longford, 1987 and 1986). Recent developments in its implementation now permit an extension of the approach to the exponential family by application of the quasi-likelihood principle (Longford, 1988b). A brief resumé of Longford's general algebraic formulation follows in the next subsection; its' full computational implications are realised in chapter 9 ¹.

¹ Footnote: Further consideration is not devoted to Goldstein (1987) simply because the software was not available at the time of analysis.

3.6.2:

Outline of the variance component model underlying VARCL

The interactive software developed by Longford as referred to in the previous subsection is referred to as VARCL (1986); it has been successfully implemented at the author's employing institution via J.A.N.E.T. (Joint Academic Network). The general model assumes that we have N units (respondents) which are each grouped into n clusters of level 2, and so on... as to define a complete (multi-level) nesting hierarchy. For instance in Longford and Aitken (1985) we have 4 levels; pupils in classrooms (level 1), classrooms as clusters of level 2 within schools (level 3) and schools within LEA's (level 4). In the applications presented in chapter 9 we have respondents within interviewer workloads (level 2) and for the PHS study measurements by year (level 1) within respondents (level 2) where interviewers are nested as factors at level 1.

For ease of exposition consider a 2-level model for respondents at level 1 and interviewers at level 2. Ignoring the second level of nesting we have an analogous situation to ordinary least squares (O.L.S.):

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \quad (3.50)$$

where an observation can be expressed in terms of two components, a systematic part and a random part, i.e. individuals are allowed to vary. Now we allow groups to vary, i.e.

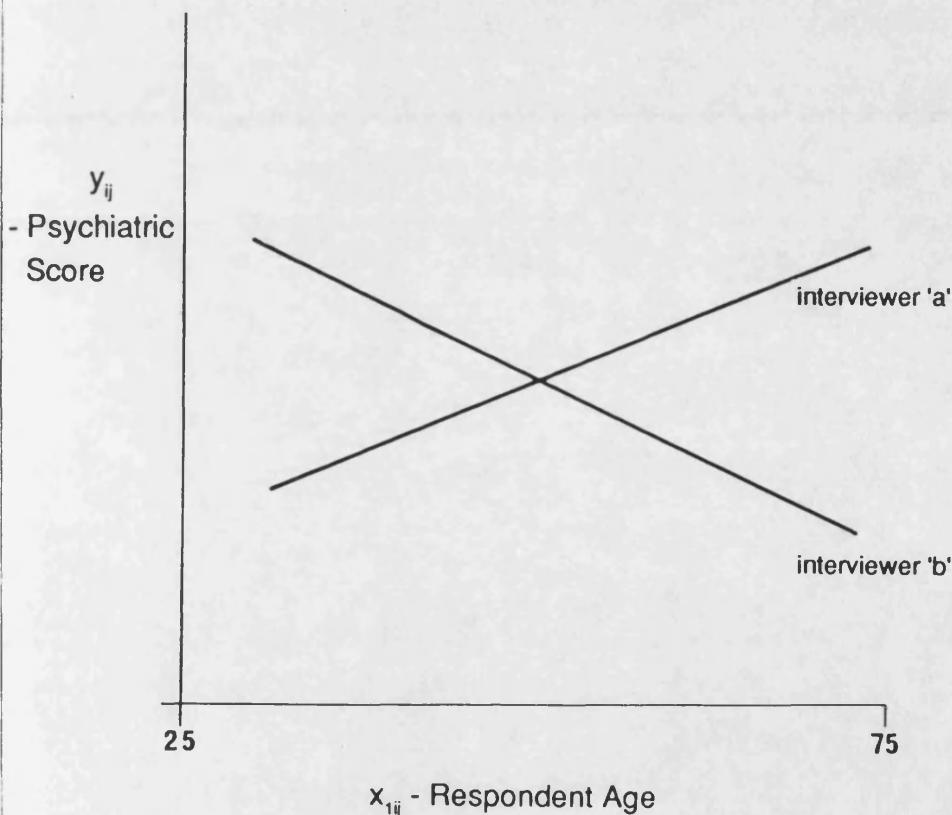
$$\begin{aligned} y &= \beta_0 + \beta_1 x_{1ij} && \text{as in the original specification} \\ &+ e_{ij} && \text{individuals now subscripted by group, } j \\ &+ \gamma_j && \text{permitting individual variation} \\ & && \text{permitting group variability} \end{aligned}$$

Formally β 's are unknown constants, e 's are independently identically distributed (i.i.d.) $N(0, \sigma_e^2)$ and γ 's are i.i.d. as $N(0, \sigma_\gamma^2)$.

This description is analogous to analysis of covariance with no interactions.

Now consider the situation where β_1 is allowed to vary, i.e. regression slopes would vary for different interviewers. For the hypothetical example in figure 3.3 below y_{ij} represents a psychiatric score for a particular respondent and x_{1ij} their respective age. The two "regression lines" represent two interviewers 'a' and 'b'.

Figure 3.3: An illustration of how regression slopes may vary in variance components analysis.



Interviewer "a" appears to obtain higher scores with increasing age of respondent. For interviewer "b" there is little variation, and, if anything, psychiatric scores tend to decline with increasing respondent age.

The model now becomes

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij} \text{ as before}$$

+ γ_{0j} with "0" subscript associated with the intercept term

+ $\gamma_{ij} x_{1ij}$ to describe random part associated with β_1

The model becomes analogous to analysis of covariance with interactions, and can be rewritten as

$$y_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{ij}) x_{1ij} + e_{ij}$$

random intercept random slope

Expanding the number of explanatory variables in the model is straightforward and leads to a general expression of the form:

$$y_{ij} = \sum_k x_{kij} (\beta_k + \epsilon_{kj}) + e_{ij} \quad (3.51)$$

for $i = 1 \dots N$ units within j units of level 2, $j = 1 \dots n$ and $k = 0 \dots p$

Note: intercept, $x_{0ij} = 1$

The $\gamma_j = (\gamma_{0j} \dots \gamma_{pj})$ form a random sample from a p variate normal distribution with mean 0 and unknown variance matrix Θ , assumed to be independent of the random sample of $e_{ij} \sim N(0, \sigma_e^2 I)$.

If the variances of Θ are constrained to zero, we have an O.L.S model. If the variance of e_0 is the only "free" parameter in Θ then we have the simple variance component model, $y_{ij} = \sum_k x_{kij} \beta_k + \gamma_j + e_{ij}$, with constant variance as introduced in subsection 3.6.1.

In general the variance of an observation is expressed as

$$\text{var}(y_{ij}) = \sigma_e^2 + x_{ij}^T \Theta x_{ij} \quad (3.52)$$

where x_{ij} is the row of the design matrix X , corresponding to the i -th observation in unit j of level 2. The implication of this formula is that the assumption of constant variance is no longer relevant, i.e. variance component models typically involve "variance heterogeneity". The model formulation can also be extended to include group (interviewer) level variables, as x_{il} , $l = 1 \dots m$, e.g. interviewer age, attitude to survey objectives or response rates.

Of course, as dimensions of variability are allowed to increase the number of covariance terms in Θ multiplies. In practice this problem is tackled by constraining most of the variances to zero. Within VARCL software, "slope by slope" covariances are treated as nuisance parameters and constrained to zero; all of the "intercept by slope" covariances are allowed to be "free" parameters subject to the constraint of non-negative definitions of Θ .

For a fuller extension of the model formulation to three and higher levels of nesting see Longford section 2, 1987. The general formulation above encompasses Searle (1970) and Rudan and Searle (1971) as a special case where the random effect at every level of nesting is contained only in the intercept term. Rudan and Searle (1971) also utilized a Fisher scoring algorithm; alternative approaches using REML and MINQUE (minimum norm quadratic unbiased estimation) for this problem are also reviewed in Giesbrecht and Burrow (1978). For a more general review see Searle, 1987¹.

¹. a MINQUE solution = a first iterate of REML. Furthermore, MINQUE demand no assumptions about the form of the distribution of y . If the usual normality assumptions are invoked, the MINQUE solution has the properties of being that unbiased quadratic form of observations which has minimum variance ie., it is a minimum variance quadratic unbiased estimator, MIVQUE.

Cross-classified random factors involve substantial computational complexity (Thompson, 1980) and are not covered by VARCL. However its application to the studies of interviewer variability described in this thesis appear to suggest merit and relevance in terms of much wider applicability. In particular it provides a vehicle to permit an interpretation of relationships between variables in a modelling context whilst allowing for the possible impact of interviewer characteristics and performance to be modelled as well. It is in this latter respect that the methodology represents an exciting development beyond that utilised in O'Muircheartaigh and Wiggins (1981), which simply allowed for the presence of an interviewer effect by inclusion of an interviewer factor in the modelling.

The concluding section of this chapter summarizes the preceding sections by embellishing Figure 3.1 to include the main points arising from the reviews. Essentially, the reader is presented with a pathway to guide their consideration of how to analyse interview effects.

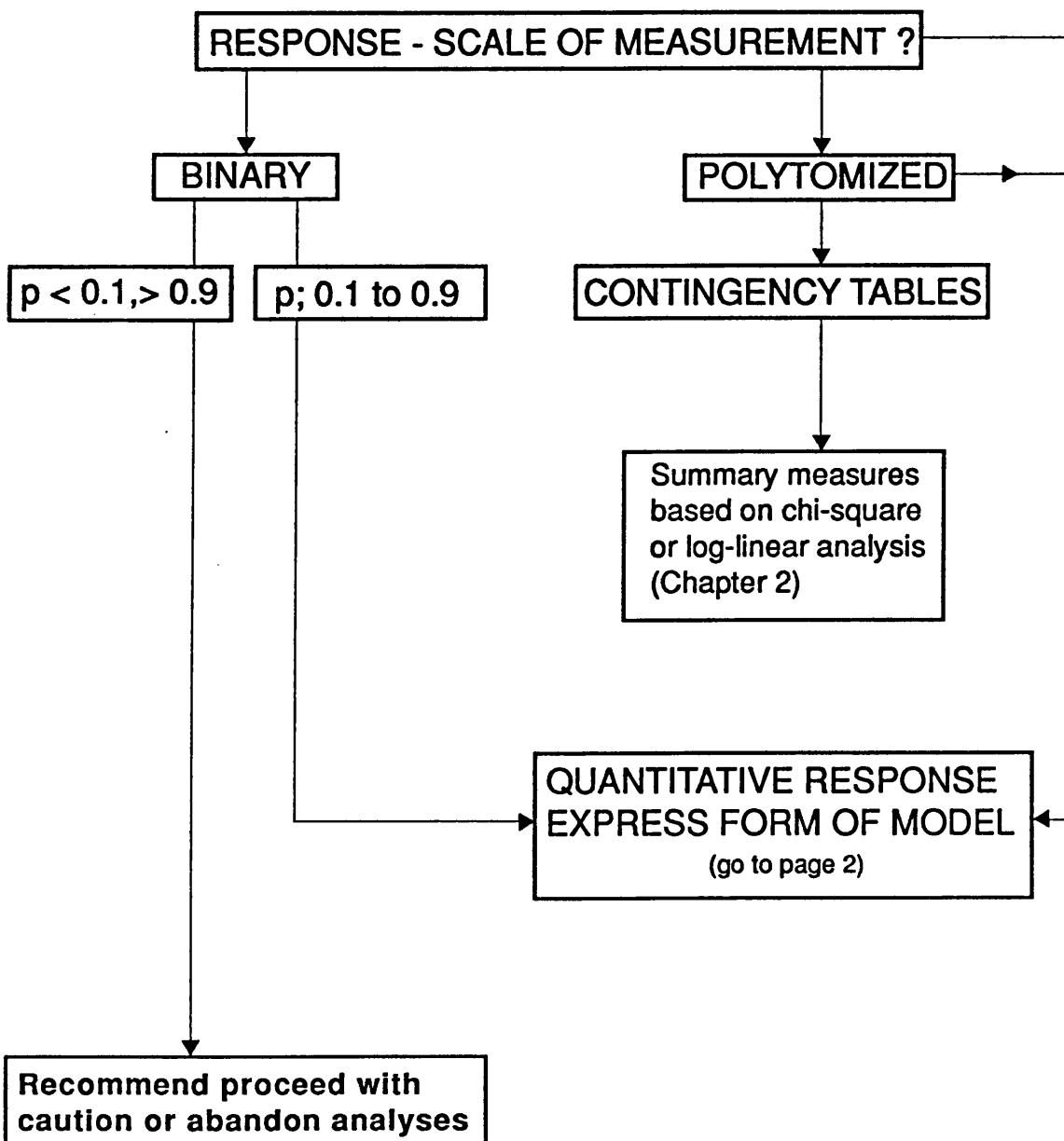
3.7: A pathway for the exploration of interviewer effects

Figure 3.4: PATHWAY FOR EXPLORATION OF INTERVIEWER EFFECTS

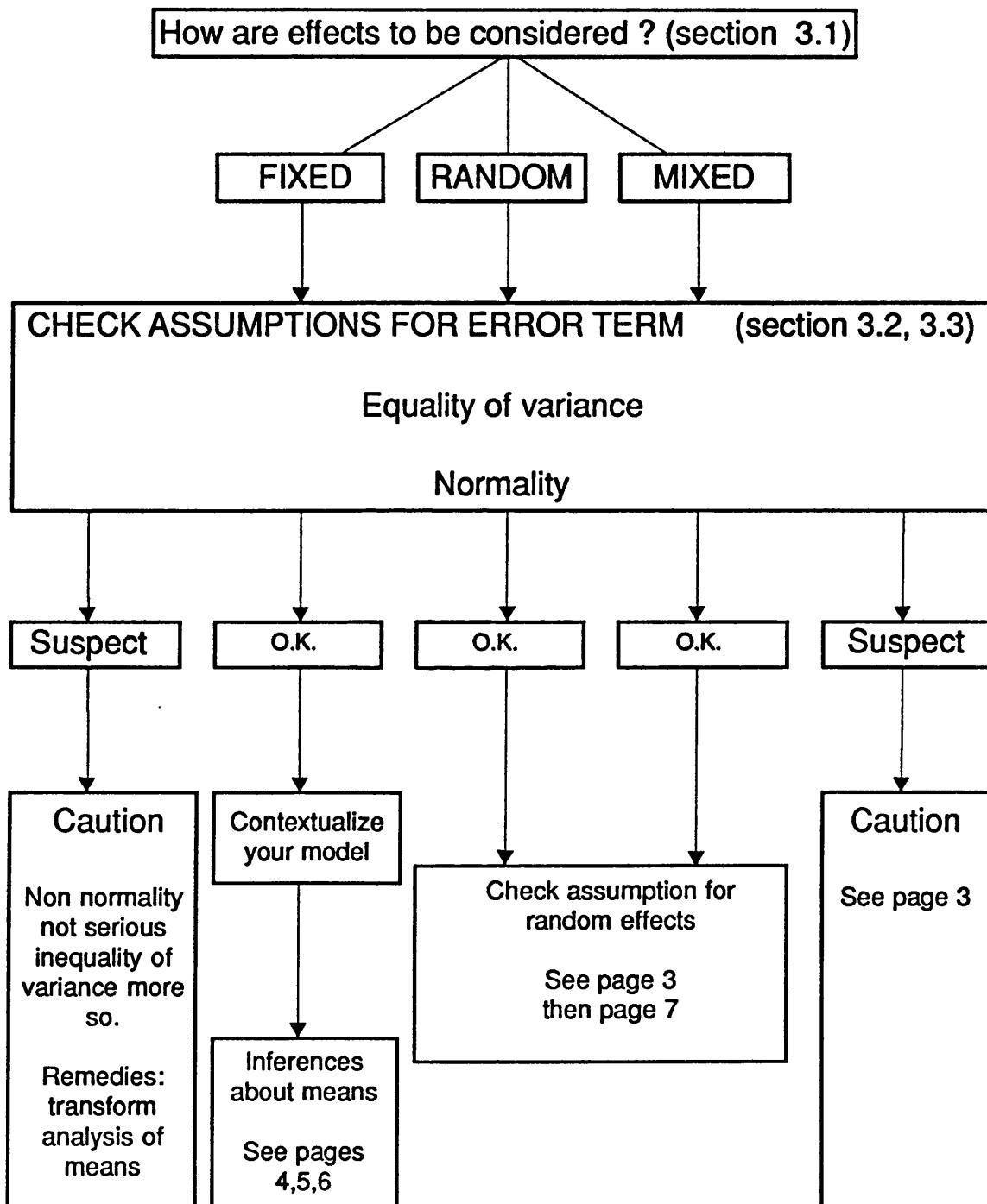
OBSERVATION AS RESPONSE single time dimension

INTERVIEWER AS FACTOR either **CROSSED** or **NESTED**

UNBALANCED DATA NORM balanced design as a special case

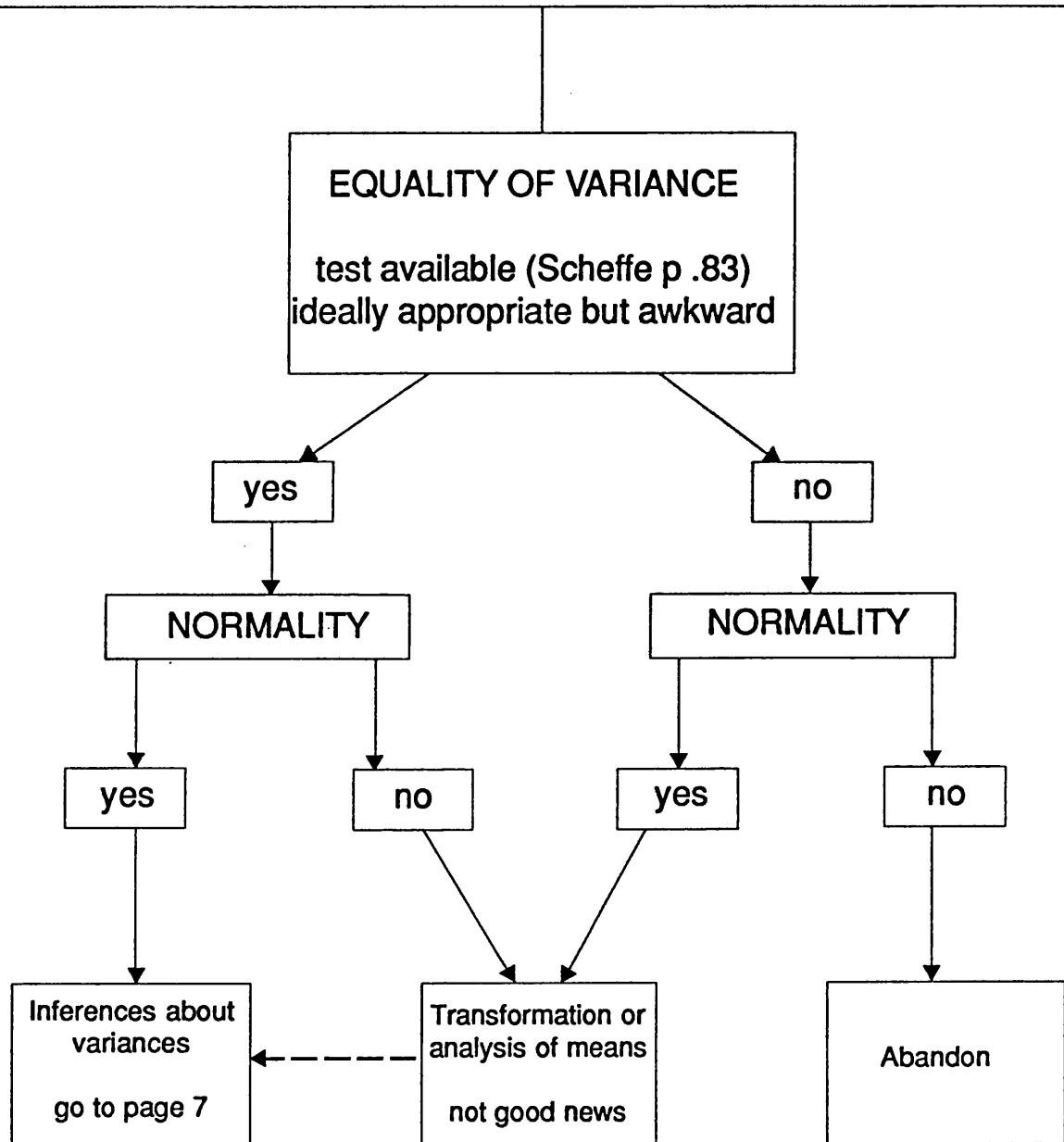


**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)



**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

CHECKING RANDOM EFFECTS ASSUMPTIONS (sections 3.2, 3.3)



**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

ONE WAY CLASSIFICATION (section 3.4)

Basic form of estimable function $\sum_{i=1}^a \lambda_i (\mu + \alpha_i)$: individual μ, α_i non estimable

b.l.u.e. $\sum_{i=1}^a \lambda_i \bar{y}_i$

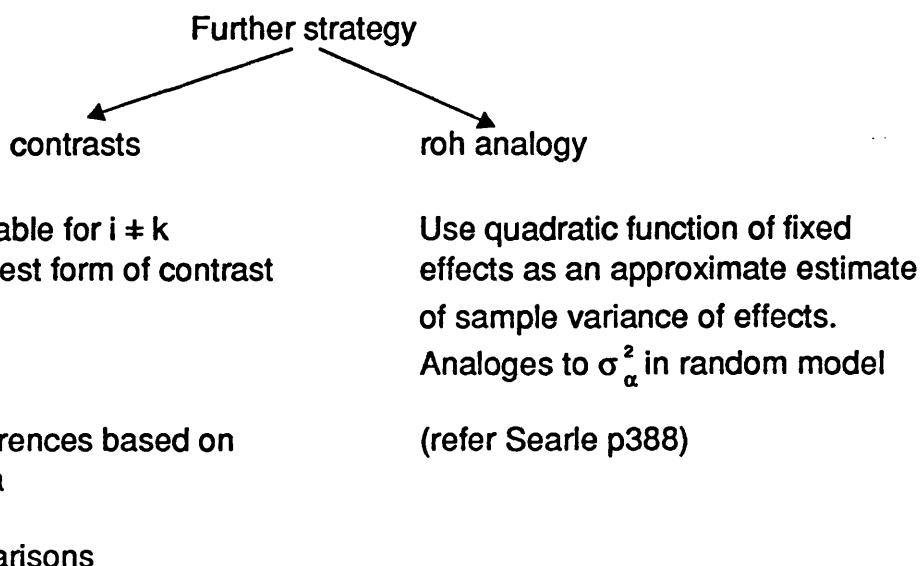
**Is model going to include
restrictions or constraints ?**

Examples of typical restrictions with implications $\sum \alpha_i = 0$ implies $\hat{\alpha}_i = \bar{y}_{i\cdot} - \sum \bar{y}_i / a$

$\sum n_i \alpha_i = 0$ implies $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{..}$

What do F - statistics imply ? $F(M)$ based on general mean tests $E(\bar{y}) = 0$

$F(Rm)$ based on model (after mean) tests
 H : all α 's equal



**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

NESTED CLASSIFICATION interviewer as nested factor (β_{ij}) (section 3.4)

Examples of estimable functions $\mu + \alpha_i + \beta_{ij}$ b.l.u.e. \bar{y}_{ij}

$\beta_{ij} - \beta_{ij}'$ for $j \neq j'$ b.l.u.e. $\bar{y}_{ij} - \bar{y}_{ij}'$

**Is model going to include
restrictions or constraints ?**

Typical examples with
implications

$$\sum_{j=1}^{b_i} w_{ij} \beta_{ij} = 0 \quad \text{with} \quad \sum_{j=1}^{b_i} w_{ij} = 1$$

Direct relevance for
interviewer effect

typically $w_{ij} = n_j/n_i$ or $1/b_i$,
no effect on testing all β 's equal. Facilitates
tests of $H : \text{all } \alpha$'s equal.

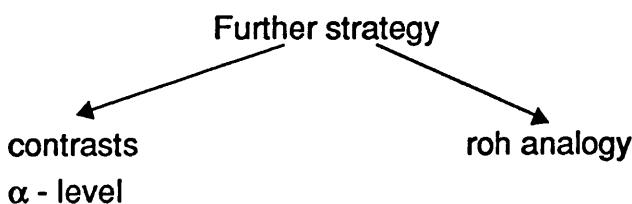
What do F-statistics imply ?
level

$F(\beta : \alpha | \mu, \alpha)$ tests all β 's equal within α
otherwise $F(M)$ tests $E(\bar{y}) = 0$
but as in Eq. (80) p256 Searle)

$$F(\alpha / \mu) \text{ tests } H : \alpha_i + \sum_{j=1}^{b_i} n_{ij} \beta_{ij} / n_i$$

$$= \alpha_i' + \sum_{j=1}^{b_i} n_{ij}'' \beta_{ij}' / n_i$$

for all $i \neq i'$: not useful unless
above restrictions imposed



$\beta_{ij} - \beta_{ij}'$ is estimable for $j \neq j'$
within area analysis could
proceed as for 1 way classification

summing sampling variance
estimates across α - levels

**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

CROSSED CLASSIFICATION (two way layout with interaction) (section 3.4)

Basic problem : two analyses of variance possible

Basic form of estimable function $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

↓
interviewer effect

b.l.u.e.

$$\bar{y}_{ij}$$

Main implication

differences between interviewer effect
non estimable without restrictions

**Is model going to include
restrictions**

where all cells filled $\sum \gamma_{ij} = 0$
enables equality of effects
to be tested; otherwise restrictions
are data dependent of form

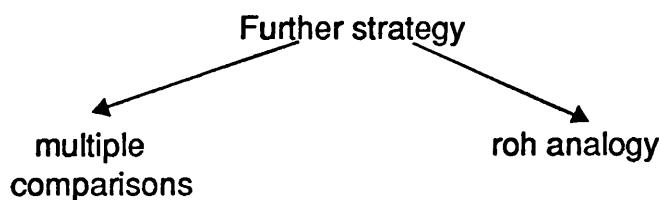
$$\sum_{j=1}^b n_{ij} (\beta_j + \gamma_{ij}) = 0$$

for estimating $\alpha_i - \alpha_i'$

F - statistics

refer Searle table 7.4
useful to check necessity of interaction
term, use $F(\gamma/\mu, \alpha, \beta)$

test on mean tests $H: E(y) = 0$ otherwise
 $F(\alpha/\mu), F(\beta/\mu),$
 $F(\beta/\mu, \alpha)$ and $F(\alpha/\mu, \beta)$ are
messy unless restrictions imposed



**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

**RANDOM OR MIXED MODELS (sections 3.5, 3.6)
VARIANCE COMPONENTS**

Advantages - directly appropriate to earlier studies of interviewer effect (in particular, Kish, 1962) where no questions posed about the random model

Disadvantages - problem of negative estimates of variance components. Variety of quadratic forms that can be used for variance estimation (refer to dual anova in fixed case for 2-way layout; which sums of squares to use? / lack of suitable criteria to choose). For unbalanced data even with normality assumptions estimators have unattractive qualities.

Estimation procedures

Analysis of variance method	not appropriate for mixed model
Henderson's method 2	designed to overcome above
Fitting constants	alternative to above
Analysis of means	O.K. where no empty cells
Symmetric sums	yet another alternative

NOTE

Fitting constants gives unbiased estimators for variance components even for mixed models - it does not provide a way of estimating the fixed effects themselves. Searle suggests two possible remedies:

- (i) use fitting constants method to estimate variance components and use these estimates in place of true estimates in generalized least squares equations.
- (ii) estimate variance components and fixed effects simultaneously under a unified procedure, such as maximum likelihood

The implication of (ii) is powerful; firstly there is no need to distinguish between fixed, random or mixed models have at least one fixed effect - the general mean - so all models can be regarded as mixed, secondly it brings into question the use of generalized least squares. All estimation can be conducted in a single unifying framework. We can also include further respondent characteristics as covariates in a modelling context.

**PATHWAY FOR EXPLORATION
OF
INTERVIEWER EFFECTS**
(continued)

ABANDON CONVENTIONAL CLASSIFICATION SCHEMES

USE μ_{ij} - MODEL or CELL MEANS MODEL (section 3.7)

Advantage - no confusion over what is an estimable function μ_{ij} - model have b.l.u.e. \bar{y}_{ij} (2 way analogue) model always of full rank, S population parameters corresponding to non-empty cells for large multi-way design may be more expeditious to use log linear analysis of cell means.

Disadvantage - onus on investigator to construct meaningful hypotheses of interest; may fall back on conventional anova parametrization as a guide.

Is model going to include restrictions not essential; presence results in smaller variance estimates for $\hat{\mu}$ important to define in terms of a prior knowledge - difficult given relative absence of application in interviewer \ effects context - hence fallback to classical framework. e.g imposing restrictions that are not always interpretable to get standard comparisons of type $\alpha_i - \alpha_i'$

Futher strategy

Comparisons of means

Methodological note: fluency in application according to Searle depends on thorough knowledge of procedures and consequent weaknesses of conventional modelling schemes.

Chapter 4:

Contextual issues: multivariate assessment and repeat measurement

Contents:

4.1 Multivariate assessment of interviewer effect

4.2 Interviewer effects in a time dimension

4.1: Multivariate assessment of interviewer effect

Once experimental design modifications are introduced into the survey process following chapter 3 the strategy for appraisal has been largely in terms of univariate assessment. Where interest is in modelling relationships between variables we have seen how the presence of interviewers as well as measures of their characteristics can be incorporated into the analysis. In this way, interviewers become part of the definition of data structure. Typically in modelling, response measures are summary indices, e.g the GHQ and annoyance scores in the ANS, or the FLP score in the PHS. These indices are summative scores based on sets or domains of individual item scores. Thus it becomes necessary to extend any analytical appraisal of interviewer effect to a multivariate context.

In such a context one question which might arise is whether or not it is possible that an index is in some sense free from any interviewer effect whilst the component parts of the summary may not be. O'Muircheartaigh (1976) extends the use of ρ to consider the issue of interviewer variance for indices.

For the general model

$$y_{hij} = y'_{hij} + \alpha_{hi} \quad (h = 1.. L; i = 1.. m; j = 1.. k) \quad (4.1)$$

where we have L items in a summative scale, then the general expression for ρ_z , to denote a ρ value for the scale score will be

$$\rho_z = \frac{V(\sum_h \alpha_{hi} / L)}{V(\sum_h \alpha_{hi} / L) + V(\sum_h y'_{hij} / L)} \quad (4.2)$$

This formulation leads to a fairly complicated expression that does not obviously provide any wisdom on the relationship between individual item values for ρ and ρ_z for the mean. By considering two reasonably restricted cases O'Muircheartaigh delivers some important implications which will be pursued in chapter 7. Firstly, by assuming constant ρ for all items in the index, then certain results follow for the relationship between the simple average of the item ρ values and the scale value defined above.

To look at individual p values may be seriously misleading, if the individual p 's are large and the average intercorrelation for individual interviewer effects across the item set (\bar{r}_α) is large then the effect on the mean may be greater than the average effect. The higher the correlations between the item scores where interviewer effects have been removed (\bar{r}'_y) the less likely that p_z will exceed p . If effects are in different directions for different items (\bar{r}_α low or negative) then the effect on the mean may be considerably less than the average p value

(p). Other particular results follow:

if $\bar{r}_\alpha = +1$ then p_z will always be greater than or equal to \bar{p}

if $\bar{r}_\alpha = -1 / (L - 1)$ then there will be no effect on the mean

Secondly an interesting case arises if one item shows much greater interviewer variability than other items in the set. In the simplest instance O'Muircheartaigh considers the situation where p is zero for all items except one. Here two general observations can be made; the larger the value of the single item only p , the more likely $p_z < \bar{p}$; the higher the correlations between the y'_{hij} , the less likely $p_z > \bar{p}$.

O' Muircheartaigh's empirical findings suggest no definitive conclusions, but it is clear that it is not sufficient to simply consider individual item sensitivity to interviewer effect. The role of the item has to be considered as well if it plays a part in defining summary indices.

The examination of several indices will be considered in chapter 7; McKennell's alpha coefficient (O'Muircheartaigh and Payne, Vol .1, 1977) will also be used to explore whether or not to 'drop' certain items from scales on the basis of the magnitude of their p -values.

Remaining within the multivariate mode of assessment we find other approaches to evaluation which concentrate on methodologies for investigating effect in terms of the structure or dimension of any underlying effect. Thus our interest is in determining to what extent different sets of interviewers are responsible for any effect across multivariable item sets. The imaginative originality of the application is due to O'Muircheartaigh (1976). Application can also be found in O'Muircheartaigh and Wiggins (1981). Primarily the approach is founded on multivariate analysis of variance; by utilizing the random effects model implicit in the univariate approach it is possible to test the hypothesis that the vectors of mean values for the L-items each interviewer obtained arose from the same L-variate normal distribution. The appropriate test statistic is Wilk's Lambda, which is the ratio of the determinant of the variance-covariance of the interviewer effects to the determinant of the variance-covariance matrix of the residuals. A multivariate analogue of the F test is used to test dimensionality of the structure (Rao, 1952). A significant result does not indicate which of the effects are non-null and on what items or combinations of items. Whilst simultaneous test procedures exist to allow some resolution of the overall significant results into rejection of detailed hypotheses on subsets of interviewers and items (see Gabriel, 1968) the author prefers the more intuitive and novel use of principal components analysis as a vehicle for disentangling structure as developed by O'Muircheartaigh (1976). Furthermore given that the use of MANOVA is subject to all of the qualifications in sections 3.1 to 3.3 for the univariate case and the unifying assumption of multivariate normality it would appear wise to simply regard the use of Wilk's lambda as a guide to the multivariate complex under study.

The use of principal component analysis relies on the existence of an L^*m matrix of interviewer effects, where individual effects (the α 's) have been estimated using a fixed effects model. The maximum number of components we can have is $m - 1$. Latent roots and consequent percentage variance explained provide confirmation of dimensionality revealed under MANOVA. By correlating component scores (which represent the "distances" between interviewer effects on a dimension) with the individual item interviewer effects it is possible to identify subsets of items which are sensitive to particular interviewer effects.

This is an attractive exploratory technique which enables us to unravel many of the subtleties of interviewer effect. O'Muircheartaigh (1976), O'Muircheartaigh and Wiggins (1981), Wiggins (1985) demonstrate how the pattern of variability between interviewers differs for different sets of interviewers for different sets of items. This evidence provides confirmation of the difficulty of categorising items with respect to interviewer distortion (Kish, 1962).

This approach underpins the analysis of item sets in chapter 7, section 7.2.2. However two points need to be made explicit in the light of the review in sections 3.2, 3.3, 3.4 and 3.5. Firstly, the distinction between random and fixed effects has not always been made clear. Both the role of the intercorrelations of item interviewer effects and responses adjusted for effect in the review of scale scores used in application of PCA make consideration of fixed effects models implicit if nothing else. Whilst blurring the distinction between fixed and random may be tolerable in the interests of exploratory analysis there is at least one lesson we should observe from earlier sections, namely the role of restrictions or constraints embedded in our linear model. For example in the one way classification would the use of $\sum \alpha_i = 0$ or $\sum n_i \alpha_i = 0$ have any substantial effect on the interpretation of structure strategy outlined by O'Muircheartaigh (1976)? This observation will be investigated in chapter 7, section 7.3.2

Secondly, the evaluations described above might suggest that one way of analysing relationships between variables in survey data analysis might be first to estimate any effect by means of an appropriate linear model and then to adjust survey responses accordingly. The view of the author is that such an imputation approach is unnecessary especially when the presence of interviewer effect is simply included in any modelling framework (refer section 3.6 and see chapter 8).

Finally, the impact of using PCA rotation techniques will be considered (section 7.3.3) and the idea of using PCA results for the first wave of the PHS to prespecify the structure of interviewer effect for the subsequent will also be presented (confirmatory factor analysis, Maxwell 1977) in section 7.3.4.

4.2: Interview effects in a time dimension

Earlier sections have assumed that we are dealing with experimental evaluations located in a single one-off time dimension. If we extend this conceptualization to include time or re-enumeration so that a survey is regarded as a single trial from among a number of possibly conceived trials under the same essential survey conditions the underlying mathematical model must shift to accommodate such a perspective.

Hansen, Hurwitz and Bershad (1961) were first responsible for providing such a framework.

The difference between an observation on the j -th unit on a particular survey or trial, t , and the expected value of that unit is called a response deviation, d_{jt} , where $d_{jt} = y_{jt} - Y_j$ where Y_j is the conditional expected value over all measurements of the element j .

Each element y_{jt} can be expressed as the sum of three components:

μ_{ij}	the individual true value
$Y_j - \mu_j$	the response bias, and
$y_{jt} - Y_j$	the response deviation.

Note that the true value only affects the bias term.

Thus an estimator for total survey error can be written

$$(\bar{y}_t - \mu) = (\bar{y}_t - \bar{Y}) + (\bar{Y} - \mu) \quad (4.4)$$

The first term consists of fluctuations about the expected value and produces total variance which in turn can be shown (O'Muircheartaigh, 1977) to consist of three components:

the response variance,
the sampling variance of the estimator (y_t)
and a covariance term involving average response deviations and sample means (\bar{d}_t , \bar{y}).

In a complete enumeration the last two terms would disappear. The response variance, can be stated as

$$\sigma^2 d_t = 1/n \sigma^2_d (1 + p(n-1)) \quad (4.5)$$

where σ_d^2 is the simple response variance ($E(d_{jt}^2)$), the effect of the variance of the individual response deviations over all possible trials, and ρ the intraclass correlation coefficient among the response deviations for a survey or trial, n is the sample size. At first glance this expression is very similar to the one introduced in section 2.2 in the context of Kish's model. To appreciate the differences it is necessary to re-introduce the idea of interpenetrating design, let's say k interviewers produce $n = mk$ interviews based on randomly assigned workloads. Correlation between response deviations is brought about by correlated interviewer effects. Then the response variance becomes $1/n \sigma_d^2 (1 + \rho(m - 1))$.

The two expressions are equivalent whenever one interviewer is responsible for all of the interviewing ($k = 1, m = n$). However one important difference remains, namely the definition of ρ . Through the introduction of time it has been possible to separate out two distinct components of variance, namely sampling variance, σ_y^2 , and simple response variance $\sigma_d^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$, which is based on a linear model formulation, $y_{ijt} = Y_j + \alpha_i + \epsilon_{ijt}$ for $t = 1 \dots T$ hypothetical repetitions of the survey. ρ is defined simply as $\sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2)$.

Without re-enumeration it is not possible to separate out σ_α^2 and σ_ϵ^2 so Kish's definition (equivalent to $\sigma_\alpha^2 / (\sigma_y^2 + \sigma_\alpha^2 + \sigma_\epsilon^2)$ above or $\sigma_\alpha^2 / (\sigma_y^2 + \sigma_\epsilon^2)$) is the only one appropriate in such a context. In general, there appears no compelling reason to use either the Hansen-Hurwitz-Bershad estimator or Kish's.

Fellegi (1964) develops a model that permits the definition of several types of correlation among the responses based on an approach that combines both interpenetration and replication. The formulation is based on the following conditions:

- (i) a simple random wr sample of $n = mk$ observations
- (ii) the sample consists of k independent subsamples, each of size m
- (iii) each subsample (S_1, \dots, S_k) is paired with another (different) subsample

- (iv) each pair of subsamples is allocated at random to k interviewers
- (v) each pair of subsamples constitute two replicates or repeat surveys for each interviewer.

Summary coefficients thus acquired are:

- (a) the correlation of response deviations obtained by the same interviewer in the same survey (δ_{2t})
- (b) the correlation of response deviations obtained by different interviewers in the same area (δ_{3t})
Common supervision or training of interviewers within trials could induce such correlation.
- (c) the correlation of response deviations obtained in the two trials; this measures recall effect (β_1)
- (d) the correlation of response deviations obtained by the same interviewer in the two trials for the same individuals (β_2)
- (e) the correlation between the sampling and response deviations for the same interviewer in the same survey. The subsample an interviewer interviews may well affect his/her attitude expectations (α_t)

Denoting simple response variance by σ_d^2 and sampling variance by σ_s^2 Fellegi also defines an index of inconsistency, I , where $I = \sigma_d^2 / (\sigma_d^2 + \sigma_s^2)$.

The Hansen-Hurwitz-Bershad p-value is close to the coefficient in (a) above, except that it involves a bias term involving " $\alpha \sigma_d \sigma_s$." If this bias term can be ignored the relationship between Fellegi's δ_z and Kish's p is

$$\delta_z I = p \quad (4.6)$$

However "I" cannot be estimated for a single survey and neither can δ_z . As O'Muircheartaigh (1977) concludes most survey designs used to assess interviewer variability do not use re-enumeration so the best comparative statistic available is p as defined by Kish.

Clearly where the opportunity to conduct repeat interviews exists Fellegi's approach will provide more information about interviewer effect than by simply replicating cross-sectional strategies. Indeed further coefficients could be added to the list above, for example β_3 , to measure the correlation between response deviations for different individuals by different interviewers in the same subsample over time. The problem encountered in estimation is that, since there are more coefficients to be estimated than there are linearly independent estimators, bias results. The best compromise solution promoted by O'Muircheartaigh is to select out the most important coefficients, where the biases are in terms of the other parameters hoping, of course, that such biases have little consequence. Whilst, in principle the experimental design for the PHS study was intended to support appraisal following Fellegi, compliance with the original design specification proved difficult, (see chapter 5). As a result small size subsamples were generated for repeat waves and such analysis was not attempted.

However time was included as a level in the variance components approach described in section 3.6. The results provided in chapter 8 conceptualize respondents as a 'level' (level 2) in a nesting hierarchy with responses nested within at level 1. Each respondent has two measurement occasions. Interviewers can be included as a 'fixed' effect for each measurement occasion.

Chapter 5: Specific survey conditions and data context for evaluations

Contents

- 5.1 **The role of illustrative studies**
- 5.2 **The Aircraft Noise Survey (ANS)**
 - 5.2.1 **Background**
 - 5.2.2 **Context for evaluation**
 - (a) interviewer effects : univariate and multivariate assessment
 - (b) interviewer effects : modelling
- 5.3 **The Physically Handicapped Survey (PHS)**
 - 5.3.1 **Background**
 - 5.3.2 **Context for evaluation**
 - (a) interviewer effects : univariate and multivariate assessment
 - (b) interviewer effects : modelling

5.1: The role of the illustrative studies

Two large scale community surveys, the Aircraft Noise survey (ANS), and the Physically Handicapped Survey (PHS) were introduced in chapter 1 as characterizing the illustrations that follow in chapters 7 and 8. These studies underpin evaluation of the impact of interviewers using some of the strategies reviewed in chapters 3 and 4. Indeed the data from these studies were generated from experimental subsamples specifically designed to investigate interviewer effect under the direction of the author. Descriptions of these studies together with some of the findings presented in this thesis will be found in O'Muicheartaigh and Wiggins (1981) for the ANS, in Wiggins (1985) for the PHS, and Wiggins et al (1990) for both the ANS and PHS data. The work on the PHS was funded under an ESRC grant, HR5971. Funds covered the additional travel costs incurred to mount a randomized allocation of twelve interviewer workloads. All interviewing was carried out by Social Community Planning Research. What follows is a brief resume of the background, description of sample structures and items used in subsequent evaluations.

5.2: The Aircraft Noise Survey

5.1.1: Background

We have seen in chapter 3 that it is only random allocation of respondents to interviewers that enables statistical methodology to identify and estimate any distortions introduced into the data by interviewers. Resultant increases in travel costs make survey researchers reluctant to undertake investigations of interviewer effect. The Aircraft Noise survey in West London provided a favourable opportunity to mount a study: a large proportion of the interviews were clustered tightly in urban areas and the randomization of interviews was not expected to increase costs unduly. It was possible to randomize the allocation of 317 addresses across eight interviewers in the study, all of which were located in the high noise stratum (Noise and Number Index contour 45). A complete description of the sample design is given by Wiggins (1980). Figure 5.1 summarizes the location study and the location of noise domains.

For the experimental sub-sample the 317 addresses issued produced 307 eligible individuals, of whom 236 (or 77%) responded. The response rates by interviewer are presented in table 5.1 below. Three of the interviewers were male and 5 were female; all had had at least 6 months but not more than 2.5 years of 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 5.1:

Response rates by interviewer

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 rates(%)	68	74	90	76	72	79	78	77	77

(From O'Muircheartaigh and Wiggins (1981))

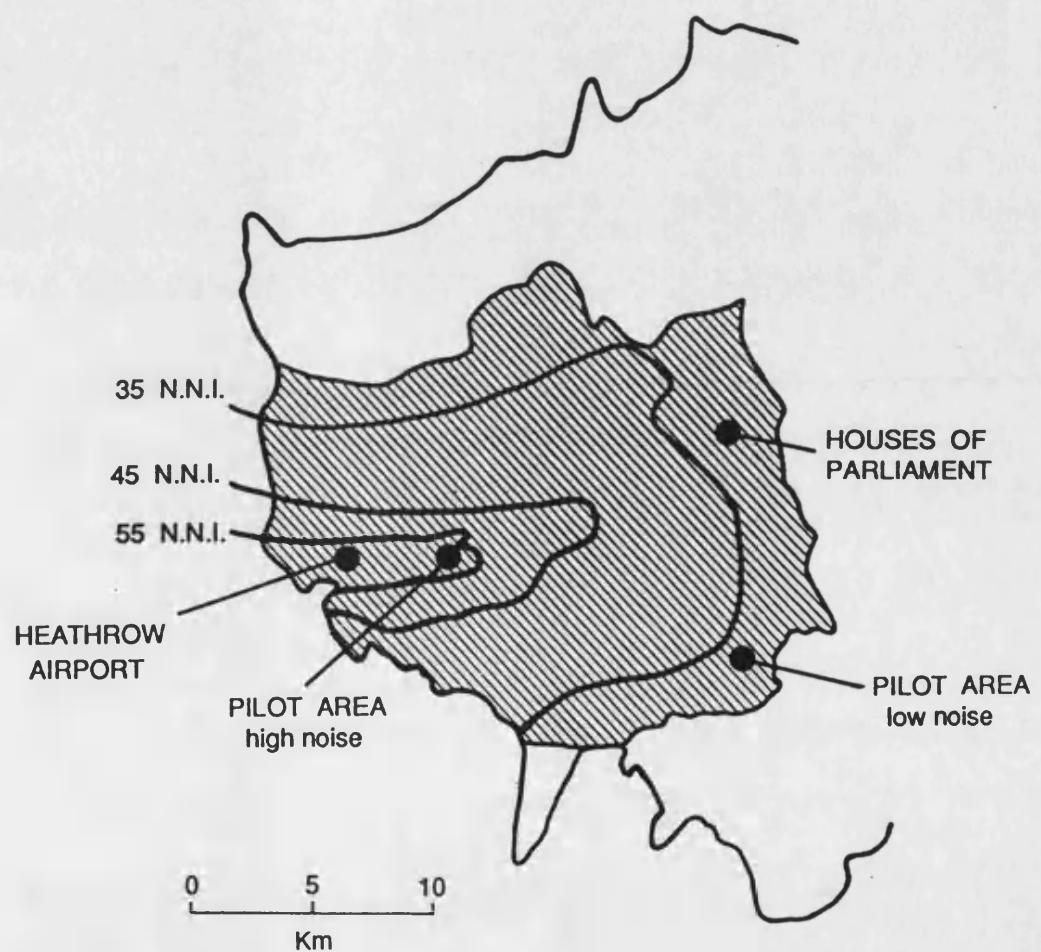
5.2.2: Context for evaluation

(a) interviewer effects : univariate and multivariate assessment

Interviewer effects were examined for three major sets of items. The question numbers below refer to their position on the main questionnaire:

- (i) Noise annoyance as measure 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
- (ii) 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)
- (iii) 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 responses for the presence of interviewer effect as 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.
Each individual item was analysed separately to estimate the magnitude of interviewer effect.
Annoyance and psychiatric morbidity scale items were analysed as a multivariate set in accord with the methodological approach reviewed in chapter 4.

Figure 5.1: West London 1977 survey of psychiatric morbidity.



(b) **interviewer effects : modelling**

The second aspect of the investigation of interviewer effects concentrates on the impact of interviewer effects on the study of relationships between items. The motivation for examining the relationships between annoyance, sensitivity and GHQ in particular arose from early work on the pilot data (Tarnopolosky et al., 1978). In that analysis the single item 'being bothered by aircraft noise' was considered as a measure of annoyance : all three items were dichotomized and considered as categorical variables to facilitate log-linear analysis. The analysis was presented in a path analytic framework following Goodman (1973), where annoyance was the main dependent variable, with GHQ acting both as a response variable and a dependent variable. A measure of noise exposure was also considered but, of course, was inappropriate in this instance as all of the interviews were concentrated in the high noise zone. In chapter 8 this early analysis was replicated, but this time the presence of an interviewer effect was also taken into account. Recent development in software applications for variance components analysis (Longford, 1986) reviewed in 3.8 enabled further insight regarding interviewer effect to be gained. The relationships described above were re-analysed so as to permit the entry of up to five interviewer level variables into the modelling. (namely : average number of calls per workload (labelled ave. calls), response rate for experimental assignment, as reported in table above (resp.rt), interviewer gender (sex), age and years of interviewing experience (yrs.exp). The total sample size is reduced to 233 to ensure the existence of complete information for all variables in the analysis. The effective workload sizes are then:

TABLE 5.2: Workloads analysed in variance component modelling

Interviewer	1	2	3	4	5	6	7	8
workload for analysis	26	28	35	29	27	30	28	30

5.3: The Physically Handicapped Survey

5.3.1: Background

Under the sponsorship of the Department of Health and Social Security (DHSS) and with the assistance of the Special Trustees of St. Thomas's Hospital, the Department of Community Medicine began in 1978 a long-term programme of research into the health and care of the physically handicapped in the London Borough of Lambeth. This research has two major objectives:

- (1) to map the course of impairment, disability and handicap in a sample study population from Lambeth households. Specifically, it was intended through repeated application of a functional limitations profile (the F.L.P. is a revised version of the Sickness Impact Profile developed by the Department of Health Services, University of Washington, U.S.A.) to relate any observed changes in dysfunction to individual, social and environmental factors. Particular attention will be given to the relationship between social structures, as measured by social class, and handicap.
- (2) to provide information for social policy development and for strategic planning decisions concerning the physically handicapped at the district, regional, and national levels. Specifically, the use and cost of services to the physically handicapped will be assessed. In addition, a comparison of perceptions of need and priorities for service as expressed by the physically handicapped with those expressed by planners and providers of the relevant services will be conducted.

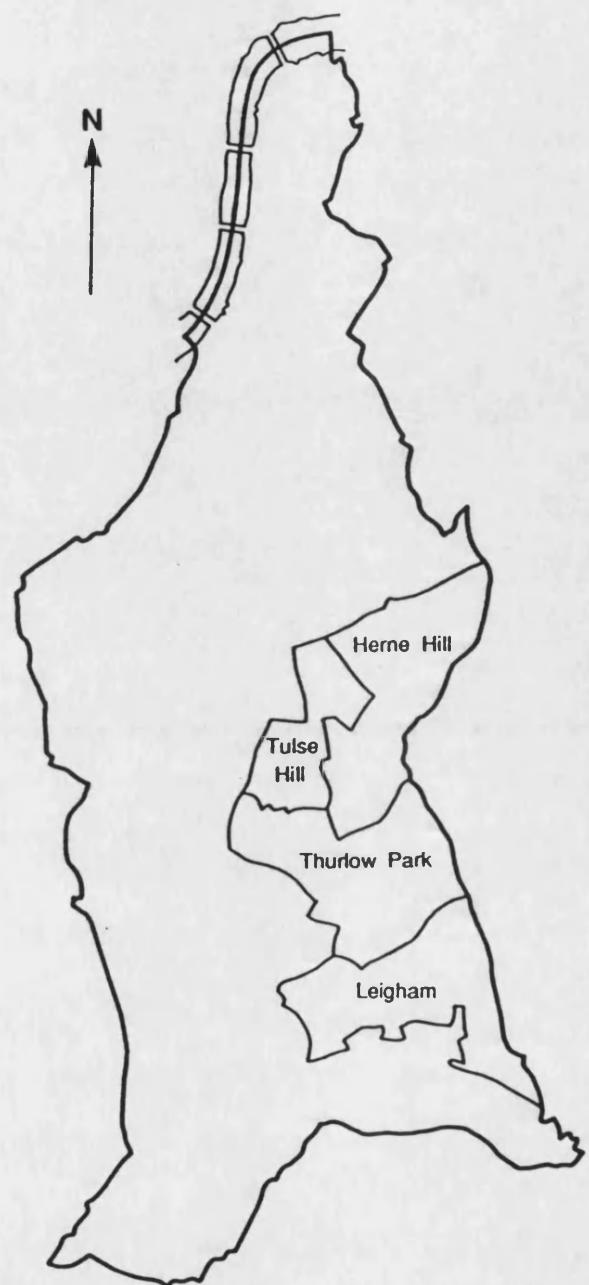
These objectives, determined in consultation with representatives of DHSS, required a large-scale, multifaceted research strategy. Over three years (1978 to 1981) five inter-related studies on impairment, disability, and handicap were conducted:

- (a) a **Screening** study to identify the physically impaired and disabled (Patrick D. et al, 1981)
- (b) a **Disability** survey to assess functional limitations, use and cost of services, and priorities for services
- (c) a **Scaling** study to measure the relative disadvantage (value) associated with functional limitations
- (d) a **Priorities** study to compare perceptions of need as expressed by the physically handicapped with those expressed by the planners and providers of services to the physically handicapped
- (e) an **Observation** study to develop understanding of the social situation of the physically handicapped and to formulate new interpretations of their problems

The opportunity to study interviewer variability arose in the context of the disability survey, although the relevance of its findings will be important not only to the survey itself but to the other inter related aspects of enquiry, particularly the scaling study and future applications of the F.L.P. The disability survey is a longitudinal survey (3 repeated interviews each one year apart) of a stratified sample of both disabled and non-disabled persons as identified in the screening study. Initially 1,100 disabled and 500 non-disabled persons were contacted; further details of the sample design are given in Patrick (1981).

As mentioned in 5.2.2 randomizing interviews will conflict with the normal field work practice of clustering workloads geographically and will, of course, increase travel costs. This constrained the scale of the study. Sufficient funds were obtained to concentrate an inquiry in the south eastern part of Lambeth, an area covering four administrative wards (Herne Hill, Tulse Hill, Thurlow Park and Leigham) containing 336 initial interviews randomly allocated across twelve interviewers. One additional factor affecting the choice of area was that it was important to be sure that interviewers would always be working alone in order to eliminate another potential source of distortion in the responses.

Figure 5.2: The Lambeth Health Survey: location of interviewer effects study:



Tables 5.3 and 5.4 give details of the original allocation scheme for both waves of the study. Table 5.5 provides further details on the success of randomization. Essentially interviewers were paired so as to facilitate an exchange of workloads for the second wave. Two interviewers, 3 and 8 were exceptions in that they would receive the same workload, and hence the same respondents, for both waves. (Such a design could permit a longitudinal analysis of the data according to Fellegi (1964))

During wave one 244 interviews were achieved in the experimental areas (73 per cent response rate); however, not all of these interviews coincided with the twelve original allocations of 28 interviews per experimental interviewer; some experimental interviewers took on part of the allocations of fellow interviewers in the experimental area. Applying the proportion of disabled identified at the screening stage of the survey as a criterion check on randomization the allocation of achieved interviews by experimental interviewer appeared to be upheld (see table 5.4). All 244 interviews were used in the analysis for wave one, though strictly only 228 were achieved according to the original allocation scheme (67.8 per cent of 336). For the second wave of the study 178 interviews were achieved by the experimental interviewers, representing 85 per cent of those interviews eligible for allocation at the end of the first wave. Unfortunately, interviewer 3 did not participate in the survey for the second wave so the analysis was based on eleven interviewers. Again some interviewers achieved interviews which were strictly outside of their original allocation but still within the experimental area. As in wave one these interviews were used in the cross-sectional analysis, though only 146 (70 per cent) complied with the original allocation scheme. Using disability status at screening as criterion check ($\hat{\rho} = 0.00$) it was decided to use all 178 interviews in the wave two cross-sectional analysis. Further details of workload allocations are provided in table 5.4.

TABLE 5.3: The experimental interviewers

(a) Allocation scheme for wave I and II

Workload allocation no:

301 302 303 304 305 306 307 308 309 310 311 312

Interviewer assigned

wave I: 1 2 3 4 5 6 7 8 9 10 11 12

Interviewer assigned

wave II: 2 1 3 5 4 7 6 8 10 9 12 11

(b) Summary of productive interviews for experiment.

Wave	Interviews issued	Interviews used for cross sectional evaluation	Interviews achieved as per allocation scheme (a) above
I	336	244 (72.6%)	228 (67.8%)
II	209	178 (85.2%)	146 (69.9%)
I & II	209	-	142 (67.9%)

(c) Effective response rates:

proportion of allocation achieved by interviewer as initially assigned to workload scheme under (a) above

Interviewer no:	1	2	3	4	5	6	7	8	9	10	11	12	Overall
	*							+	x				
Response rate I:	.71	.57	.68	.71	.71	.71	.71	.71	.46	.78	.71	.64	.68
II:	.81	.20		.95	.75	.65	.75	.70	.68	.77	.50	.95	.70

* not initially part of experiment; entered late in fieldwork on wave I, illness in wave II meant reallocation necessary

+ the only interviewer to repeat interviews with same respondents

x not initially part of experiment; entered late on in wave I

**TABLE 5.4: Completed interviews by disability status
(defined at screening)**

Interviewer	wave 1					wave 2		
	Dis- abled	Non Dis- abled	Propn. dis.	Int. Ach.	Dis- abled	Non- disabled	Propn. dis.	Ints. Ach.
1	18	2	.90	20	9	4	.69	13
2	10	6	.63	16	6	2	.75	8
3	17	11	.61	28	-	-	-	-
4	16	11	.59	27	28	17	.62	45
5	12	8	.60	20	9	6	.60	15
6	15	5	.75	20	11	2	.85	13
7	16	4	.80	20	12	3	.80	15
8	16	4	.80	20	12	2	.86	14
9	9	4	.69	13	12	3	.80	15
10	16	6	.73	22	7	3	.70	10
11	16	4	.80	20	6	3	.66	9
12	12	6	.67	18	15	6	.71	21
TOTALS	173	71	.71	244	127	51	.71	178

$\hat{\rho} = .00$

$\hat{\rho} = 0.00$

**TABLE 5.5: Workload character:
univariate analysis of selected
socio-demographic variables**

variable	^ p	P level
Respondent's sex (024)	- .03	.901
Respondent's age (025)	.02	.162
Respondent's marital status (026)	- .02	.841
Respondent's work status (027)	- .02	.825
Nos. in household (868)	- .02	.845
Respondent's colour (070)	.02	.162
Respondent's birth place (074)	.01	.293

Note: original variable number in brackets

k = 20.33

Clearly the practicalities of ensuring rigorous compliance with experimental design specifications is a serious problem for field controllers who may not share the same objectives as the survey methodologist. In this experiment due to the contractual nature of the field work the researcher had to abdicate most of the day to day management of the experiment to field supervisors who had to put completion of fieldwork as their major aim. It was decided for subsequent variance component analysis to use only those interviews which complied with the original design specification.

The evaluation presented in Chapter 7 sets out to summarize the extent of interviewer effects on individual items, category scores and multivariate item sets (specifically the F.L.P.) over repeat interviews one year apart. The influence of various social, demographic and attitudinal variables related to the interviewers themselves on the magnitude of interviewer effect is also considered. Replication of the analyses permits a valuable check on item sensitivity. "Time" is also introduced as a dimension in modelling relationships.

5.3.2 Context for evaluation

(a) interviewer effects : univariate and multivariate assessment

Interviewer effect is considered wholly in terms of the Functional Limitations Profile. The profile consists of 135 items subdivided into twelve major sets or domains describing possible restrictions or limitations in daily living. Table 5.6 summarizes the twelve categories.

TABLE 5.6: Functional Limitations Categories for Disability Survey*

<u>Category</u>	<u>Limitations or Dysfunction</u>
(SI)	Social interaction
(A)	Ambulation or locomotion activity
(SR)	Sleep and rest activity
(E)	Eating activity
(W)	Usual daily work
(HM)	Household management
(M)	Mobility and confinement
(BCM)	Body care and movement
(C)	Communication activity
(RP)	Recreation and pastime activity
(AB)	Alertness behaviour
(EM)	Emotional behaviour

* Adapted from the Sickness Impact Profile developed by Department of Health Services, University of Washington, Seattle, Washington, U.S.A.

Each domain contains a number of individual items, for example, "I sit around half asleep" from the sleep and rest domain. The respondent is asked to say whether the statement describes them to-day and if so is this description due to their health. Each item is treated as dichotomous. Additionally utility/judgement weights can be assigned to each item if they are positively endorsed ("yes" it describes respondent to-day and it is due to their health) in order to define 12 individual domain scores (as summative indices) or a global index (the "F.L.P") to describe a respondent's level of dysfunction. The exact wording of each item is listed in Wiggins (1985) and the appendix.

Multivariate assessment focusses on the behaviours of items within the twelve domains presented in table above.

(b) interviewer effect : modelling

Data for modelling relationships between items in the presence of interviewer effects was guided by analysis reported by Charlton (1981). The final selection of respondent characteristics was clearly not meant to be definitive. The intention is simply to provide further useful illustration of the potential of variance component analyses in the context of understanding interviewer variability.

The global F.L.P score (as a proportion of the maximum score attainable * 1000) was used as a response variable throughout (for further detail on refinements to this measure as a response see Charlton et al., 1983).

The original modelling strategy adopted by the PHS researchers focussed on those respondents aged 25 to 75, where age was categorized into three separate subgroups for modelling purposes. Insufficient numbers in the repeat waves for the experimental subgroup analysed here meant that respondent age has been simply defined as a continuous explanatory variable (age). Other variables included for explanatory candidature are sex (level 2=female), working status (work, level 2=inactive), self assessment of health, a 5-point rating defined as a continuous variable to reflect use of "formal" services, here attendance at hospital outpatients in the two weeks prior to interview (level 2=yes,label hsp.out).

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. Unfortunately it was not possible to obtain precise information on interviewer response rates in the experimental area as some interviewers were reallocated parts of assignments originally intended for other experimental interviewers and also completed work in non-experimental areas. This item was excluded from subsequent analyses.

Rather than simply repeat modelling strategies separately for each wave data was first structured to allow an assessment of the longitudinal character of the study to take place. This requirement together with the need to ensure the presence of complete information on all items implied substantial reduction in the effective sample for analysis. Firstly, two interviewers were excluded because they had reduced workloads in one of the waves; interview 2 was ill during wave 2, interviewer 9 entered the study late in wave 1. Secondly interviewer 3 was not originally part of the group selected to participate in the experiment and dropped out before the second wave. The implication of these events for longitudinal analysis meant that some respondents only had one observation in terms of the original allocation scheme described in table 5.3(a). Such workloads were dropped from the analysis (thus losing interviewers 1 and 10 as well). Altogether **seven** interviewer workloads were used for combined level analysis. Table 5.7 presents a summary of the response sets analysed. Data was also analysed separately for each wave.

All of the interviewers included in the experiment were "panel" interviewers which implies they had a longstanding commitment to the agency. Table 5.8 represents complete data on interviewer characteristics.

TABLE 5.7: Subsamples analysed for combined and separate wave analyses in the Physically Handicapped experiment

interviewers by original pair	respondents wave one	wave two
4	15	14
5	14	15
6	11	8
7	8	11
11	16	8
12	8	16
8*	12	12
Totals	84	84

*only interviewer to see the same respondents in both years

Note: for the analyses presented in chapter 8 interviewers have to be numbered consecutively. The conversion is as follows:

4	becomes	6	(9600)
5	"	1	(8090)
6	"	2	(9240)
7	"	3	(9260)
11	"	7	(9905)
12	"	5	(9560)
8	"	4	(9390)

TABLE 5.8: Interviewer Characteristics

<u>Identification</u>	<u>Effective R/R</u>		<u>ATDP/ SCORES</u>			<u>SEX</u>	<u>AGE</u>	<u>PANEL EXPERIENCE</u>	
	<u>I</u>	<u>II</u>	<u>I</u>	<u>II</u>	<u>PANEL EXPERIENCE</u>			<u>a lot</u>	<u>quite a lot</u>
1	.71	.81	54	56	F	57	Yes	a lot	
2	.57*	.20*	68	64	F	-	Yes	a lot	
3	.68	- *	-	-					
4	.71	.95	87	71	M	54	Yes	a lot	
5	.71	.75	78	84	F	49	Yes	a lot	
6	.71	.65	89	84	F	53	Yes	a lot	
7	.71	.75	81	82	F	42	Yes	quite a lot	
8	.71	.70	82	81	F	51	Yes	quite a lot	
9	.46*	.68	92	99	F	47	Yes	a lot	
10	.78	.77	74	71	M	29	Yes	quite a lot	
11	.71	.50	66	63	M	27	Yes	some	
12	.64	.95	86	80	F	43	Yes	quite a lot	
	.68	.70	.77.9	76.2	Propn. (F)	All panel Mean		Propn. " alot"	
overall (11)	overall (11)			=.73 (11)		age = 45.2(10)		= .55 (11) = .33 (3)M = .63 (8)F	
								= 36.7 (3)M = 48.9 (7)F	

* not originally in expt.
or drop out through illness

Chapter 6: Resume of evaluation strategy

Contents

- 6.1: Overview**
- 6.2 Individual item sensitivity and extensions of methodologies to appraise the impact of roh (p)**
- 6.3 Multivariate assessment**
- 6.4 Modelling**

6.1: Overview

The previous chapter describes the context for the evaluation that follows in chapters 7 and 8. Both the ANS and PHS studies embrace modifications in their design to premit the appraisal of interviewer effect. Readers are reminded that the intention of chapters 3 and 4 is to inform the empirical appraisals that follow. In particular, it might be useful to consult the summary pathway outlining the theoretical consequences of experimental design (section 3.9) before reading chapter 7.

As there are so many items in both surveys only selected variables will be used to demonstrate the methodologies used in the appraisals.

Chapter 7 begins with univariate item sensitivity to interviewer effect (sections 7.1 and 7.2) as well as presenting variance estimates for individual values of roh . The methodology underlying this achievement is described in the next section. Additionally, the consequences of replicating appraisals for the PHS is considered. Constraints on relative subsample sizes and design realization did not pemit Fellegi's methodology (1964) to be implemented. Multivariate assessment of item sets described in chapter 5 is contained in section 7.3, as well as the consequences of mis-specifying constraints contingent on the experimental design.

Chapter 8 is devoted solely to the exploration of the impact of interviewer effect in the context of modelling relationships between variables. Modelling software, GLIM (see Nelder and Baker, 1978) and VARCL, (see Longford, 1978a) is used for this purpose. Variance components analysis allows the analyst to consider not only the presence of interviewers on a response, but the influence of the interviewer characteristics as well. VARCL is also used to explore the passage of time on the response patterns. Essentially, for the PHS data each respondent defines a level in the hierarchy with two observations (one for each year) nested within the individual.

6.2: Individual item sensitivity and extensions of methodologies to appriase the impact of roh

All of the items described in chapter 5 are first assessed in terms of Kish's roh criterion based on one way classifications of unbalanced data under random effects assumptions for interviewer effect. As negative estimates of variance components arise directly as a result of sampling variation negative values of roh are reported as opposed to the typical convention of equating such values to zero. Sample proportions for binary items are reported in tables 7.8, 7.9 and the appendix. Estimates of sampling variability are rarely, if ever, included in the literature despite the existence of sound methodology for the estimation of variance. To redress this gap jack-knife repeated replication methodology as described in Frankel (1971), Kish and Frankel (1970,1974) Krewski, Rao (1981) and Wolter (1985) has been applied for selected items.

In particular variance estimates are obtained by direct application of the jack-knife method described by Wolter (1985), (section 4.2, p154-156). We assume that the sample can be partitioned into k separate groups, where each group corresponds to an individual interviewer's workload of responses. Let $\hat{\rho}_{-i}$ be an estimator of roh but computed from a reduced sample obtained by omitting the i -th interviewer's responses.

A pseudo-value of roh is defined as:

$$\hat{s\rho}(i) = \hat{\rho} - (k-1) \hat{\rho}_{-i} \quad (6.1)$$

A jack-knife estimator of the variance of ρ is then:

$$\text{var}(\hat{\rho}) = \frac{1}{k(k-1)} \sum_{i=1}^k (\hat{s\rho}(i) - \hat{\rho})^2 \quad (6.2)$$

A further innovation to assess the behaviour of individual values of ρ for separate items comprising a multivariate set is to plot cumulative percentages distributions for these item sets against theoretically derived distributions of roh based on $F-1/(F-1+k)$ (Kish, 1962), where k denotes the average workload.

Values of F can be generated by using a NAG algorithm (G04AEF, (1987). When combined with relevant values of k theoretically constructed roh values can be obtained. In this way a theoretical cumulative distribution of roh can be created to compare with any corresponding empirical distribution.

Thus for any values of ρ and k one can always obtain corresponding values of F and α levels of probability or significance. For table 6.1 (containing roh values for 46 items) reproduced from Kish (1962) with $a=20$ interviews and $n=440$ interviewers and an average workload $n/a=k=22$ the following approximate values of α, F and ρ are obtained.

TABLE 6.1: statistical evaluation of ρ

α	.05	.10	.25	.975
F	1.65	1.45	1.20	.46
ρ	.03	.02	.01	-.025

(taken from Kish, 1962)

Similarly for the Physically Handicapped study wave one with $k=21$ we obtain:

α	.05	.10	.25	.975
F	1.83	1.59	.69	.34
ρ	.04	.03	-.01	-.03

Considering the impact of interviewer effect on individual items Kish (1962) poses the question "what can we learn on the basis of this data?" He argues that to use conventional levels of significance such as $\alpha = .01$ or $\alpha = .05$ would generally be wrong for rejecting the null hypothesis of zero interviewer effect. He suggests criterion like $\alpha = .25$ or even $\alpha = .50$ as operational decision rules if one considers the null hypothesis as doubtful a priori and when the cost (or risk) due to acceptance is high. Using $\alpha = .25$ translates into p 's of .01 for table 6.1, indicating that about half of the items in that study might be suspected of showing some interviewer effect.

Clearly, deviations of expected cumulative relative frequency distributions of roh rest on the assumption of the null hypothesis being true. This might be too naive for Kish.

Section 7.2 also illustrates of the use of alternative summary measures of interviewer effect utilizing the practice of Gales and Kendall in 1957 and subsequent contingency table summaries as reviewed in Everitt (1977). The relationship between individual values of roh and the roh value for indices based on subsets of such items is explored using O'Muircheartaigh (1976).

6.3 Multivariate assessments

Univariate analyses provides information on the sensitivity of individual items to interviewer effect. In order to know more about the way in which different subsets of items are affected by interviewers simultaneously multivariate analyses of variance is initially conducted to indicate the strength of an underlying interviewer component. If present, the nature of interviewer distortion is explored by use of principal component analyses applied to matrices of interviewer effects. These matrices are constructed by first estimating the net interviewer biases-the set of α 's in the appropriate one way classification model described in 3.6 - and arranging the

estimates by each item so defining an interviewer by item effects matrix. For the ANS the annoyance scale responses are considered as an 8 by 10 matrix of interviewer effects; for the PHS the twelve F.L.P domains are analysed separately for each wave with 12 interviewers in wave 1 and 11 interviewers in wave 2. The implication resulting from lack of careful attention to constraints imposed in order to obtain estimates of effect are illustrated together with a consideration of the practice of rotating dimensions (components to identify subsets of items which may be sensitive to particular interviewers. The consequence of this practice is also considered as to which interviewers are deemed to predominate in any influence. The consistency of interviewer influence over time is also considered utilizing "confirmatory factor analysis" as described by Maxwell (1972), and implemented using LISREL (refer to Joreskog and Sorbom, 1986).

6.4 Modelling

Modelling relationships between variables in the presence of interviewer effect is first illustrated in chapter 8 using data from the ANS by means of log-linear analysis using GLIM and reported in O'Muircheartaigh and Wiggins (1981). Analysis is then replicated using VARCL and illustrations deepened to allow for the presence of interviewer characteristics both for the ANS and the PHS. In combined analysis for the two waves in the PHS "time" enters the modelling as a level of nesting. The application of VARCL implies two considerations, firstly, a necessary reconceptualization of the data structure and secondly, an appreciation of recent software developments.

Whilst recognition of clustering effects in sampling and design dates back to around the 1950's, e.g. (Cochran, 1953) the lack of suitable computational algorithms has led to the proliferation of the "design effect" approach-resulting at best in an adjustment technique for subsequent parameter estimates (regression coefficients). The obvious attraction of flexible variance component algorithms is that "clustering effects" can be dynamically accounted for in a modelling context. Structure or hierarchy in such contexts is defined to take a better account of reality than the "flat" single level assumption

underlying data generation in the O.L.S instance. In the three cross-sectional experiments that follow all hierarchies are defined as two levels where respondents are nested within interviewers.

At first it may appear that the introduction of repeat interviews of the PHS would simply define a third level in the hierarchy for time or year of interview. Unfortunately this simple extension of the structure does not apply. Different interviewers are not present within each year, nor are different groups of respondents nested within each interviewer as the diagram below would imply.

Figure 6.1:

A three level hierarchy for time and interviewer

level 3: year

1 2 3

level 2: interviewer within year

1 2 3 4 a

level 1: respondent within interviewer

1 2 3 4 k

A way round the strict definition of 'hierarchy' is to consider observations (one for each year) as nested within respondent. In this sense, the respondent becomes the second level in the hierarchy with observations for both years as the first level. The impact of the interviewer can be explored by including a 'factor' to define presence. Figure 6.2 summarizes the hierarchies used to conceptualize the modelling illustrated that follows in chapter 8.

Figure 6.2: Resume of hierarchies used in variance component modelling in chapter 8

Cross-sectional analyses (ANS and PHS year one and two)

level 2: interviewer

1 • k

level 1: respondents

1.....n_k

Combined level analysis (PHS only)

level 2: respondent

•
1 2 k, where k = 84

yr 1 yr 2

level 1: observation, one for each year, ie. n_k = 2

A major attraction of variance component modelling is that variables can be defined at every level in the design, enabling the analyst to make an assessment of the variability between groups at each level in the hierarchy. There are distinct parts to any model specification-the "fixed" part and the "random" part. The fixed part is a description of the relationship (regression) of the response in terms of the explanatory variables for the "average" or "typical" interviewer. The random part is a description of the variability of this relationship among the interviewers and provides a "quality check" on the stability of the corresponding fixed effect. Conditional means or "residual" effects of the random effects associated with the groups of every level can also be obtained. As illustrated in Aitkin and Longford (1986) these conditional means are useful diagnostic agents and in the applications that follow provide a way of exploring how the relationship of a random effect (e.g respondent's age) varies with the response given other fixed effects currently in a model (the conditional aspect) for each group (typically interviewer). In a fully interpenetrating design (Mahalanobis, 1946) geographical areas would define a possible third level in the hierarchy, though none of the examples meet such description. Current software implementation only handles up to three levels of nesting. Appropriate error distributions can be selected to include Normal, Binomial, Poisson or Gamma specification; interaction effects can be defined simply as well as transformations. For full details on the software power the reader is referred to VARCL user manual (Longford, 1988a). The package was transferred to the author's site (VAX-VMS) via the Joint Academic Network (J.A.N.E.T)

CHAPTER 7:

Empirical evaluations of the pattern of interviewer effect

7.1 Univariate analyses: the Aircraft Noise Survey

7.1.1 Estimates of roh for individual items

7.1.2 Results of variance estimation for roh on selected items

7.1.3 An illustration of an alternative assessment strategy for categorical items

7.2 More univariate analyses: the Physically Handicapped Survey

7.2.1 Estimates of roh for individual items across time

7.2.2 An illustration of the behaviour of category scores

7.3 Multivariate assessment of selected item sets

7.3.1 Multivariate assessment of the Functional Limitations Profile

7.3.2 An illustration of mis-specifying constraints when estimating interviewer effect

7.3.3 A review of the practice of rotating axes when determining the pattern of interviewer effect

7.3.4 A strategy for confirmatory factor analysis

7.4 Summary of main findings

7.1:

Univariate analyses: the aircraft noise survey

It should be stressed that values of ρ in the following subsections are estimates of underlying parameter values based on one way random effects models for unbalanced data. Using procedures described in chapter 6 estimates of sampling variability are included for selected items. For items comprising multivariate sets, namely the annoyance and GHQ scales for the ANS study and the functional limitations profile for the PHS study, theoretical distributions of cumulative distribution functions are provided for comparison with observed empirical distributions.

7.1:

Estimates of ρ for individual items

In all, 41 individual items were analysed to estimate the magnitude of the interviewer effect. Table 7.1 presents the distribution of the values of ρ (the estimated interviewer effect) over all the items, together with the value of the multiplier $[1 + \rho (k-1)]$ for each value of ρ .

TABLE 7.1:

Distributions of values of ρ for 41 questionnaire items

ρ	No.	Cumulated No.	Value of $[1 + \rho (k-1)]$
		Items	of Items
-.02	3	3	.42
.00	8	11	1.00
.01	4	15	1.29
.02	6	21	1.58
.03	7	28	1.87
.04	4	32	2.16
.05	3	35	2.45
.06	2	37	2.74
.07	2	39	3.03
.09	1	40	3.61
.10	1	41	3.90

the distribution of ρ values provides strong evidence of interviewer effect. More than a quarter of the items show values of ρ significant at the 0.05 level and eight items have values of ρ 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 $\rho = 0.035$ is to double the variance of that item. Thus if we were to estimate the variance of such an item in the usual way, we would underestimate the true variance by 50%, with serious consequences for significance tests based on that item.

Figures 7.1 and 7.2 present the data in graphical form, distinguishing between the items in the annoyance scale and the GHQ.

Figure 7.1:

**Relative cumulative frequency distribution
for $\hat{\rho}$ on 10 Annoyance scale items.**

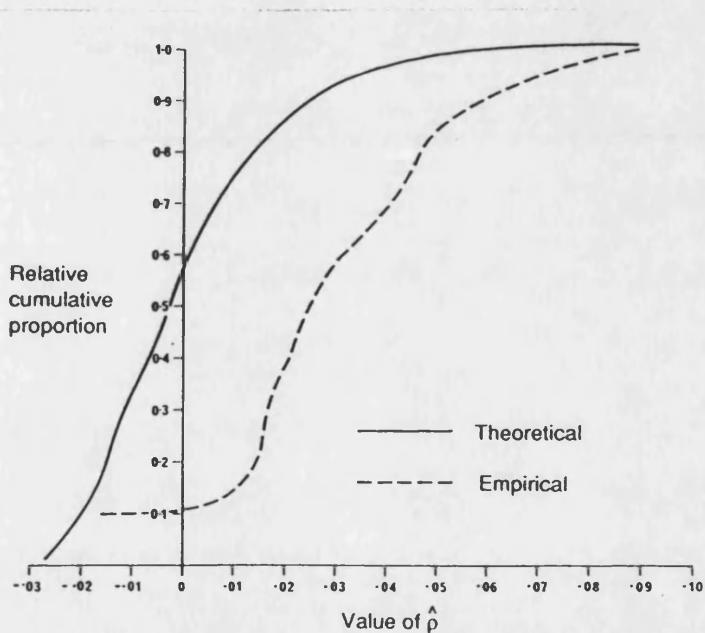
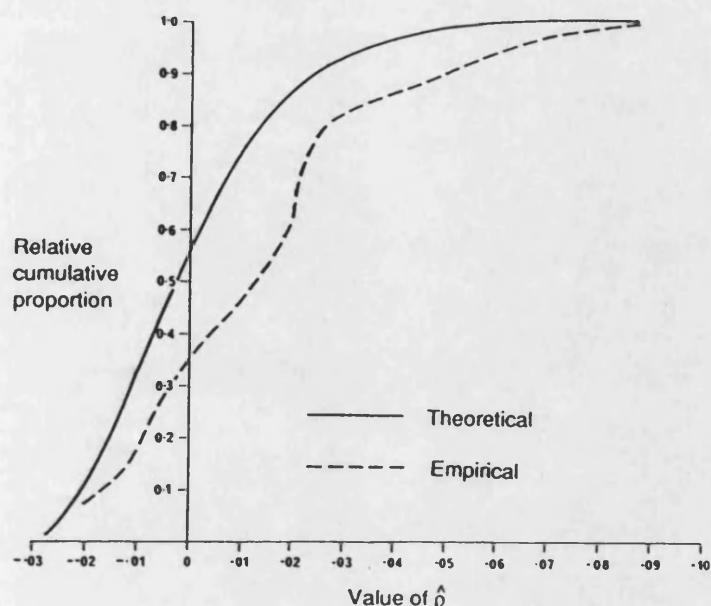


Figure 7.2:

Relative cumulative frequency distribution for \hat{p} on 29 GHQ scale items.



From the diagrams 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. The relationship between the p-values for the individual items comprising that scale has been investigated by O'Muircheartaigh (1976), who established that a consideration of the items individually is not sufficient to determine the sensitivity of the scale score to interviewer effect. The p-values for the scale scores for each of the two scales has been compared. These, together with estimates of the corresponding multiplier effect, and standard error are given in Table 7.2.

TABLE 7.2:

Values of \hat{p} , appropriate standard error and variance multiplier for the GHQ and annoyance scale scores

Scale	\hat{p}	s.e(\hat{p})	k	Multiplier $1 + \hat{p}(k-1)$
GHQ total score	.0599	.0511	29.5	2.77
Annoyance score	.0256	.0370	29.5	1.59

The results in table 7.2 are surprising. 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. In fact, 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 observation can be confirmed by a multivariate assessment of the interviewer by item "effects" matrix as reported in O'Muircheartaigh and Wiggins (1981).

7.1.2:

Variance estimation for roh on selected items

In the previous sub-section, estimates for the standard error of ρ were presented for both the GHQ - score and the annoyance scale. The methodology for their estimation was outlined in section 6.2. Noticeably standard error estimates are almost as big, if not bigger (as the case of the Annoyance score) than the original sample estimates of ρ . With a small number of degrees of freedom (7) these results may not be surprising and may throw further caution on any assessment. The author feels it provides a strong case for the routine production of variance estimates in interviewer variance evaluations. Table 7.3 illustrates the estimation methodology for the GHQ score. It is also another useful way to gain an insight into interviewer influence; estimates of ρ_{oh} vary between .0135 and .0784 depending on which interviewer workload is excluded. The lower limit .0135 is obtained when interviewer 3 is excluded and represents a dramatic 77% reduction in the value of ρ_{oh} . Clearly the way in which interviewer 3 handled the presentation of the GHQ, or indeed further aspects of this person's interviewing style warrants more investigation.

TABLE 7.3:

**An illustration of variance estimation for
the GHQ score**

Overall value of ρ_{oh} = .0599

		<u>ρ_{oh} estimate</u>	<u>pseudo- value</u>
Based on all interviewers except	1	.0504	.1260
Based on all interviewers except	2	.0702	- .0124
Based on all interviewers except	3	.0135	. 384
Based on all interviewers except	4	.0579	.0735
Based on all interviewers except	5	.0784	- .0703
Based on all interviewers except	6	.0688	- .0028
Based on all interviewers except	7	.0739	- .0387
Based on all interviewers except	8	.0552	.0924

NOTE: Pseudo (i) = $\{k \times \hat{p} - (k - i) \hat{p} \text{ excluding interviewer } i\}$

$$\text{Variance } (\hat{p}) = \frac{1}{k(k-1)} \sum_{i=1}^k \{ \text{pseudo } (i) - \hat{p} \}^2$$

where $k = 8$

$$= .002616$$

$$\text{Standard } (\hat{p}) = .0511$$

error

7.1.3:

An illustration of an alternative assessment strategy for categorical items

The Aircraft Noise Survey is typical of many community surveys in that a lot of items are categorical. Whilst analysis of variance procedures may be reasonable for binary responses (see 2.5), their unthinking use for polytomized items may be misleading. There are a wealth of appropriate summary statistics for contingency tables containing categorical items (Everitt, 1977). What follows is a brief illustration of their application for univariate assessment of interviewer effect.

Functions of chi-square in the context of interviewer variability were first suggested by Gales and Kendall (1957) to compare pairs of interviewers across polytomous response categories. In table 7.4 below the response patterns for the single item "altogether how much are you bothered by aircraft noise" has been dichotomised and presented by interviewer.

TABLE 7.4:
Degree of expressed annoyance with aircraft noise by interviewer

Annoyance	Interviewer							
	1	2	3	4	5	6	7	8
"little to none"								
0-3	3	14	17	6	15	11	3	11
-----	-----	-----	-----	-----	-----	-----	-----	-----
"a lot"								
4+	23	14	19	23	13	20	25	19
-----	-----	-----	-----	-----	-----	-----	-----	-----
totals	26	28	36	29	28	31	28	30

chi-square = 25.84, 7 d.f., p <.001

The table demonstrates a highly significant finding : with ample evidence to demonstrate extremes in response patterns between interviewers. As the standard chi-square statistic depends on sample size use of Pearson's coefficient of contingency (1904) or Cramer's coefficient (1946) could be considered, (see Everitt, 1977 for details). Here respectively 0.31 and 0.34. Both coefficients always lie between 0 and 1 and attain a lower limit of zero in the case of complete independence (or no interviewer effect). Cramer's coefficient is preferred as it has more desirable properties regarding attainment of the upper limit 1 in the case of complete association. The standard errors of all of these coefficients can be deduced from the standard error of the chi-square statistic, and the formulae is given in Kendall and Stuart (Vol. 2, ch 33, 1961).

As an alternative to simply considering functions of chi-square to explore the presence of interviewer effect it is possible to assess a range of possible log linear models for the data in table 7.4. The results of fitting such models is presented in table 7.5, each model has a number of independent parameters associated with its specification and a goodness of fit statistic (approximately chi-square, see Payne C, 1977). Model 1 is the simplest that could be fitted to a two way table and would suggest that an observation is equally likely to fall into any cell; thus for an I X J table, the expected frequency = $n/IJ = F_{ij}$. This corresponds to the log - linear model $\log F_{ij} = \mu$. Models 2 and 3 have one of the variable effects excluded, eg. "interviewer null" describes a situation where interviewers have no effect on the distribution of frequencies, so that categories of annoyance are equiprobable within interviewer. Model 4 corresponds to the conventional hypothesis underlying the chi-square calculation. Inspection of the goodness of fit statistics reveals that none of the four models provide a satisfactory "fit", so the only possible solution would be to consider a fifth model which would include an interaction between interviewer and annoyance. This model is called the saturated model because it includes all possible terms. Such an outcome provides clear evidence for interviewer effect ie. different respondents respond quite differently on annoyance for different interviewers.

TABLE 7.5:**Log-linear models for two-way table of
annoyance by interviewer**

Model	Chi-square	d.f	P-level
1. Null	54.99	15	0.0000
2. Interviewer null	30.07	14	0.0075
3. Annoyance null	52.89	8	0.0000
4. Interviewer annoyance independent	27.98	7	0.0002

7.2:

More univariate analyses: the physically handicapped survey

The mode of presentation is similar to that used in sub section 7.1.2, the major scale item being the F.L.P. Results are, of course, replicated across two occasions, wave one and wave two respectively, facilitating additional insights about interviewer behaviour over time. In the appraisal that follows particular F.L.P categories, namely **Sleep and rest** and **Recreation** will be used to illustrate methodologies.

7.2.1:

Estimates of ρ for individual items across time

Table 7.6 represents a percentage cumulative frequency distribution for individual item ρ values obtained in both waves. Around 17% of the items have a ρ value of .05 or above in both waves (resulting in an underestimation of variance of nearly 50% if interviewer effect is ignored). A similar finding to one reported by Collins (1978) in a study of disability in Southampton. Following on from table 7.6, figures 7.3 and 7.4 summarize the cumulative distribution frequencies of $\hat{\rho}$ for the observed and theoretically derived distributions.

TABLE 7.6:
Univariate analysis of FLP items

<u>^ p value</u>	Cumulative per cent of items		Multiplier values on sample variance	
	<u>Wave 1</u>	<u>Wave 2</u>	<u>k=21</u>	<u>k=16</u>
-.05	1.5	3.0	-	-
-.01	4.0	14.0	-	-
.00	41.0	47.0	1.00	1.00
.02	58.0	63.0	1.40	1.30
.03	70.0	75.0	1.60	1.45
.05	84.0	83.0	2.00	1.75
.08	95.0	93.0	2.60	2.20
.21	100.0	100.0	5.20	4.15

(.155 , .212)
maximum maximum
in wave 1 in wave 2

Figure 7.3:

Relative cumulative frequency distributions
for $\hat{\rho}$ on 135 FLP scale items. (Wave 1)

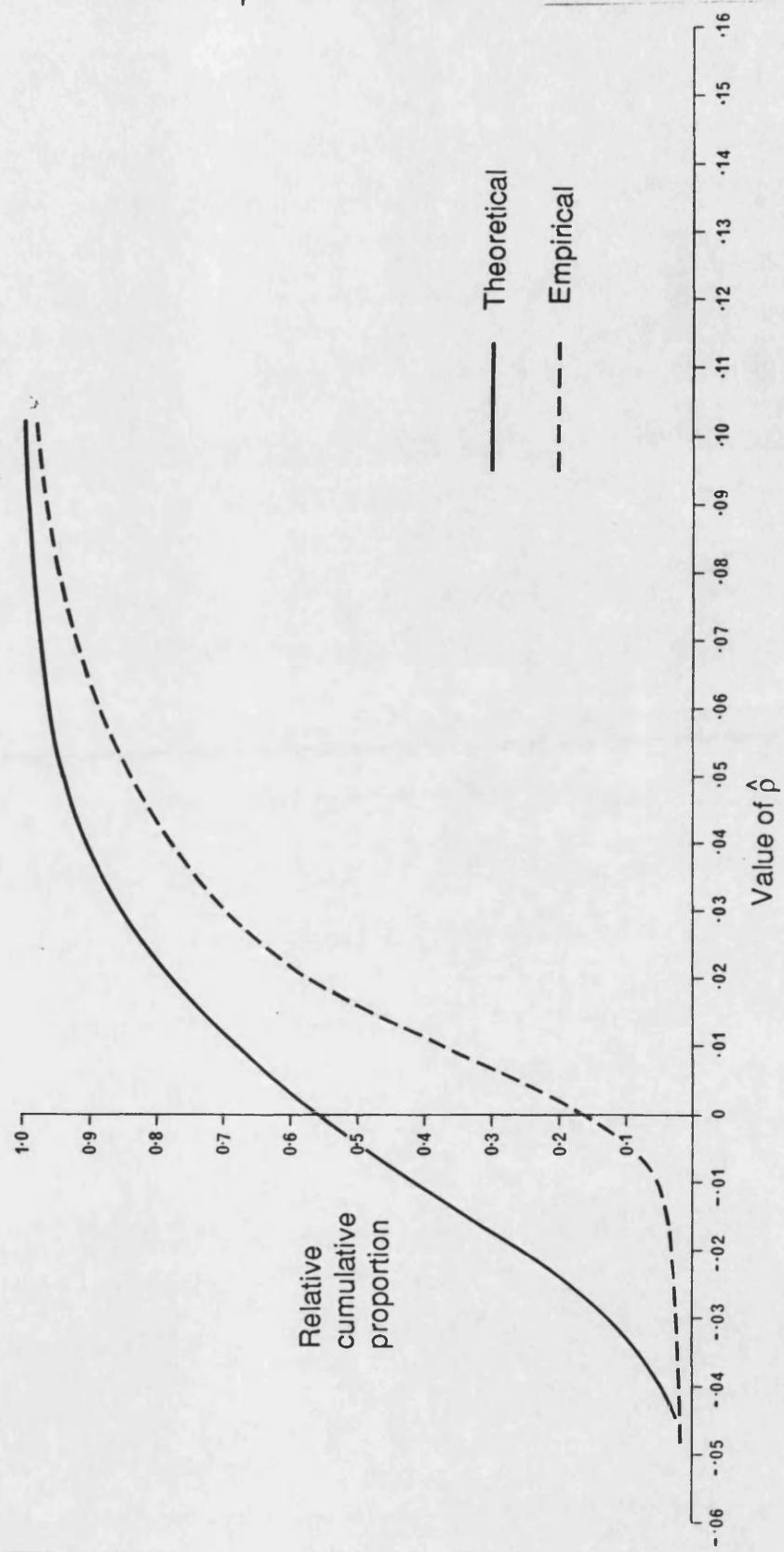
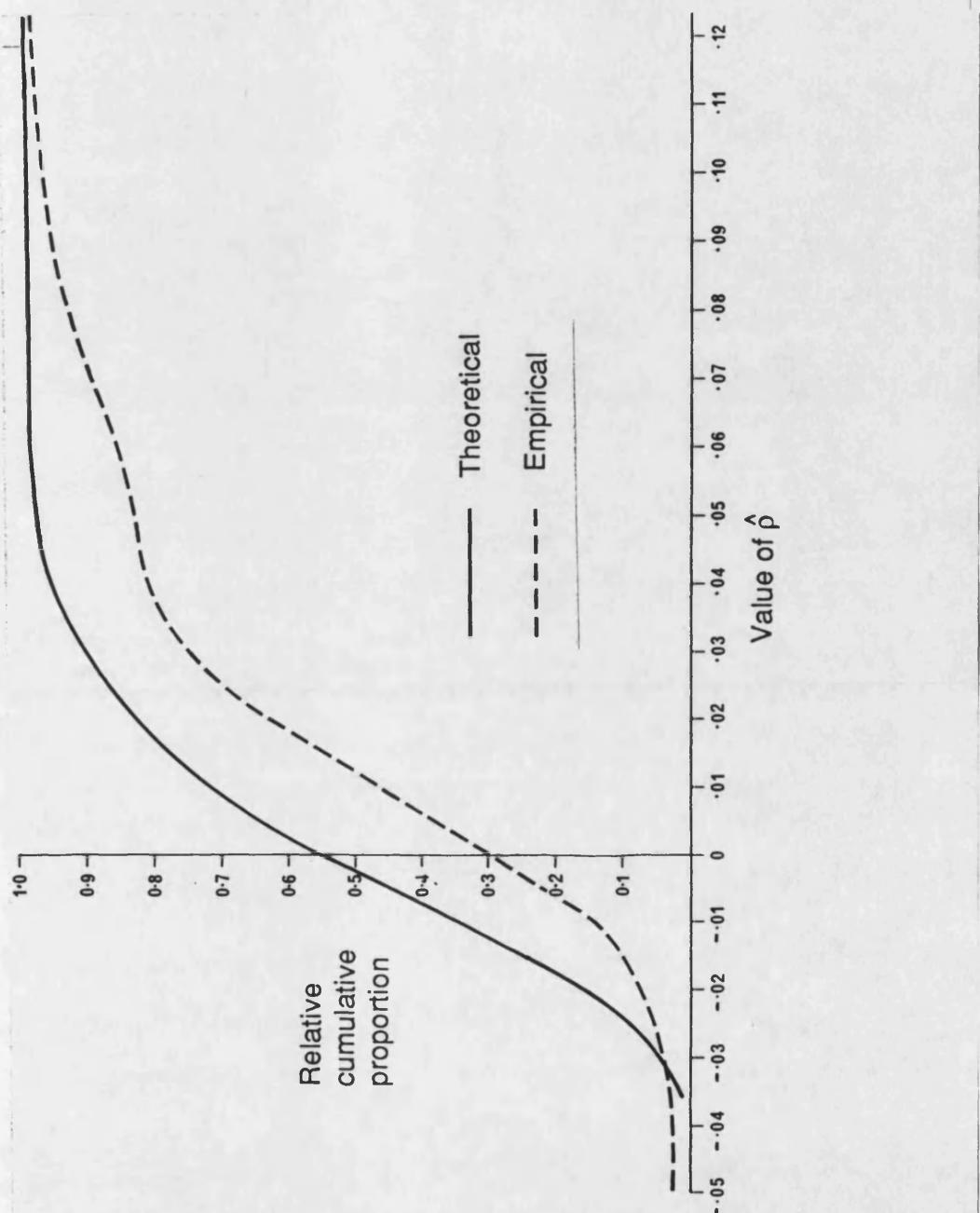


Figure 7.4:

Relative cumulative frequency distributions
for \hat{p} on 135 FLP scale items.



In both years individual items appear more sensitive to an effect than one might expect by chance alone. If anything items appear to exhibit slightly more variability in wave one; this is confirmed to some extent when examining the \hat{p} values for the F.L.P overall (wave 1 : .031, st. error = .056; wave 2 : - .013, st. error = .032). Large standard errors reflect considerable interviewer volatility both within and between waves. The initial similarity in the global results for individual \hat{p} values soon disappears when an inspection of intercorrelations between values commences. Across all of the scale items the correlation is .19, which only suggests modest consistency of "effect" overtime. Examination of intercorrelation within each F.L.P category reveals more discrepancy. Table 7.7 summarizes the findings.

TABLE 7.7:

**Intercorrelations of \hat{p} values for F.L.P
items between waves 1 and 2**

<u>category</u>		<u>correlation</u>
Eating	(9 items)	- .34
Bodycare	(23 items)	.42
Ambulation	(12 items)	.48
Mobility	(10 items)	- .26
Work	(8 items)	- .01
Household		
Management	(10 items)	- .20
Recreation	(8 items)	- .23
Sleep & Rest	(7 items)	.30
Communication	(9 items)	- .06
Alertness	(10 items)	.25
Emotion	(10 items)	.45
Social	(20 items)	.16
Interaction		
ALL ITEMS F.L.P	(135 ITEMS)	.19

There is clearly a wide range of inter correlation, from -.34 for Eating items to .48 for Ambulation items. Thus, whilst global inspection of the magnitude of ρ_{oh} values suggests item sensitivity to interviewer it is more difficult to disentangle which items are consistently sensitive over time. Indeed one is tempted to suggest that different items are sensitive to an effect on different occasions. Four categories, namely Bodycare, Ambulation, Sleep and Rest and Emotion appear to generate consistent evidence for interviewer behaviour over time. Table 7.8 reproduces all of the univariate information for Sleep and rest. Sample estimates for individuals who endorse each statement and say 'it is due to their health' are included to provide a caution on the wisdom of analysis of variance methodology for binary items. In this sense conclusions for two items, 325 and 327 may be unreliable.

TABLE 7.8:

**Univariate assessment for items
comprising the Sleep and rest scale**

ITEM	roh estimates and standard errors			sample estimate of proportion endorsing item		
	<u>WAVE 1</u>		<u>WAVE 2</u>		<u>WAVE 1</u>	<u>WAVE 2</u>
	\hat{p}	$se(\hat{p})$	\hat{p}	$se(\hat{p})$		
323	28	35	- 9	19	.07	.03
324	25	38	-30	12	.25	.24
325	14	19	- 5	23	.06	.02
326	32	26	-13	29	.13	.11
327	28	40	207	200	.05	.03
328	71	44	62	54	.23	.19
329	13	19	14	40	.13	.09

Note: Correlation of roh values between waves = .30

All \hat{p} values multiplied by 1000.

Key to items comprising "Sleep and rest" items

- 323 I spend much of the day lying down to rest
- 324 I sit during much of the day
- 325 I am sleeping or dozing most of the time - day & night
- 326 I lie down more often during the day to rest
- 327 I sit around half asleep
- 328 I sleep less at night, for example, I wake up too easily, I don't fall asleep for a long time, or I keep waking up
- 329 I sleep or nap more during the day

Average values of roh for the two waves are similar, .030 and .025 respectively. Individual items do exhibit variation in the magnitude of any effect despite a correlation of .30 overall. If anything items 328 and 329 appear to be accounting for most of the association. It would seem reasonable to reappraise the wording for items 328 and 329 and to probe more carefully to discover how or why other items are more sensititve in one wave compared to the other.

In contrast, Recreation, provides a negative intercorrelation (-.23) across the two waves. However, closer inspection of individual item roh values is warranted. Table 7.9 details the information, average roh values are similarly high in both waves (.047 and .041 respectively).

TABLE 7.9:

**Univariate assessment of items
comprising the "recreation" scale**

ITEM	roh estimated and standard error		sample estimate of proportion endorsing item		WAVE1	WAVE2
	WAVE	WAVE 2	WAVE1	WAVE2		
	\hat{p}	$se(\hat{p})$	\hat{p}	$se(\hat{p})$		
315	29	49	78	44	.22	.23
316	41	45	-24	16	.32	.30
317	34	44	103	53	.11	.07
318	134	155	28	33	.05	.03
319	9	42	44	48	.22	.19
320	50	33	21	32	.25	.21
321	38	30	15	42	.30	.26
322	32	76	62	52	.25	.19

Note: Correlation of roh values between waves = -.23

All \hat{p} values multiplied by 1000.

Key to items comprising "recreation" items

- 315 I spend shorter periods of time on my hobbies and recreation
- 316 I go out to enjoy myself less often
- 317 I am cutting down on some of my usual activities
- 318 I am not doing any of my usual inactive pastimes, for example, I do not watch TV, playcards, or read
- 319 I am doing more inactive pastimes instead of my other usual activities
- 320 I take part in fewer community activities
- 321 I am cutting down on some of my usual physical recreation or more active pastimes
- 322 I am not doing any of my usual physical recreation or more active pastimes

Although average ρ values are similar in both waves the negative intercorrelation between individual values would seem to reflect large shifts in the numerical magnitude of ρ for particular items. From these illustrations it would appear that global appraisal of item sets or subsets is not in itself sufficient. It may be a useful clue to consistency (or lack of) over time but must be accompanied by a careful examination of individual item information¹. Conversely, individual item appraisal is not enough without referencing those items to the summary scale scores they generate. Each F.L.P category is used to define a summated score (the proportion of items positively endorsed). Charlton (1981) has shown that there is good association between an F.L.P category score using judgement weightings and the total number of items positively endorsed. It is, therefore, of great practical interest to pursue an appraisal of item sets to explore whether or not category "scores" could be free of interviewer effect (individual effects "cancelling out" in some sense) when individual items are not free of an effect.

7.2.2:

An illustration of the behaviour of category scores

O'Muircheartaigh (1976) first explored the relationship between the average of a set of ρ values within a scale and the ρ_z values for their respective category mean. He found the relationship to vary, averaging the ρ 's neither consistently overestimates or underestimates ρ_z . Table 7.10 below presents a similar interpretation for the twelve F.L.P categories across both waves. However, close inspection of the table does reveal some interesting observations about the behaviour of ρ_z .

¹ Footnote: full details of univariate assessment for the remaining ten F.L.P categories is in the appendix.

TABLE 7.10:

**Interviewer effect for F.L.P category
means**

Category	No of Items	Average \hat{p}		\hat{p} for Category Mean (\hat{p}_z)	
		I	II	I	II
Eating	9	26	9	22	21
Body Care	23	15	24	10	24
Ambulation	12	16	3	-2	-11
Mobility	10	25	4	-2	-22
Work	8	-36	12	21* ¹	-28* ²
Household mgmt	10	11	-2	26	-30
Recreation	8	70	41	64	27
Sleep and Rest	7	36	40	47	11
Communication	9	-3	25	-4	54
Alertness	10	19	24	11	49
Emotion	9	51	24	9	7
Overall	135	20	18	31	-13
Inflation Factor		1.4	1.28	1.60	-

(all \hat{p} values multiplied by 1000)

*₁ much reduced k (9.33)
all other scales k = 20.33

*₂ much reduced k (6.9)
all other scales k= 16.2

Firstly, about half of the scale scores have reasonably low \hat{p}_z values across both waves, namely Body Care, Ambulation, Mobility, Work and Social Interaction. This implies that whilst some individual \hat{p} values are quite high within these categories the summary indices appear to be quite robust with regard to

interviewer effect. Of the remaining categories the evidence appears less conclusive, though Sleep and Rest and Recreation appear consistently sensitive to interviewer effect.

An interesting opportunity presented by replicate data was to invoke a methodology based on McKennell (1977) to look for ways of reducing the number of items within a scale on the basis of high individual p values in order to lower the p_z value obtained on the resulting category mean. This methodology was originally applied to look for good internal consistency for item responses within a scale as a vehicle for set construction. McKennell's strategy is based on inter-item correlations rather than p -values to select a subset of items from a pool. The coefficient " α " is used as a reliability criteria to discard items, where

$$\alpha = \frac{m \bar{r}_{ij}}{1 + (m-1)\bar{r}_{ij}} \quad (7.1)$$

where m = the number of separate items under consideration ; and \bar{r}_{ij} = the average of all of the interitem correlations.

Alpha will be greatest when \bar{r}_{ij} is a maximum. Each item contributes to \bar{r}_{ij} according to its average correlation with the other items in the pool, labelled \bar{r}_{ij}^* . Items are ordered in terms of \bar{r}_{ij}^* and this gives the order in which they should be discarded to preserve the maximum possible \bar{r}_{ij} and alpha values for remaining items. Table 7.11 overleaf illustrates the procedure for 9 attitudinal items taken from McKennell (1974).

TABLE 7.11:

Reliability after discarding successive items (from A.C. McKennell, Surveying Attitude Structures, pp. 45-55, Elsevier, Amsterdam, 1974)

item no.	8	7	3	1	9	6	2	4	5
m	9	8	7	6	5	4	3	2	-
\bar{r}_{ij}^*	0.18	0.26	0.27	0.26	0.31	0.33	0.33	0.36	0.40
\bar{r}_{ij}	0.30	0.33	0.35	0.37	0.40	0.41	0.42	0.50	-
Alpha	0.79	0.79	0.79	0.76	0.77	0.73	0.69	0.67	-

m = number of items remaining in the scale after item on the left discarded
 \bar{r}_{ij}^* = initial correlation of each item with the other eight items
 \bar{r}_{ij} = average intercorrelation of the m items
 Alpha = reliability value for the m items

The approach has been applied to two F.L.P scales with consistently high \hat{p}_z values, namely Recreation and Sleep and Rest. Items are first ranked in descending order of individual \hat{p} values - initially using wave 1 results on wave 2 data as one might do in practice and then, in the interests of symmetry, using wave 2 results on wave 1 data. Each time an item is discarded the resulting \hat{p}_z value is calculated. Obviously the desired objective would be to minimise \hat{p}_z in a manner that results in a consistent subset of items being obtained. Table 7.12 summarizes the results for "Sleep and Rest".

TABLE 7.12:

**Using individual \hat{p} values as a criterion for
eliminating items from a category scale:
Sleep and rest**

(a) Using wave 1 results on wave 2

Items in (descending) rank order of \hat{p} value	328	323	324	327	329	326	325
No of items in scale	7	6	5	4	3	2	-
\hat{p} values (x 1000) obtained for wave 2	11	2	0	32	23	10	-

(b) Using wave 2 results on wave 1

Items in (descending) rank order of \hat{p} value	327	328	329	324	325	323	326
No of items in scale	7	6	5	4	3	2	-
\hat{p} value (x 1000) obtained for wave 1	47	50	44	51	44	42	-
NOTE: average \hat{p} value x 1000		I		II			
Range	16-92		0-207				

Applying wave 1 results to second wave data appears to suggest that excluding item 328 "I sleep less at night, for example, I wake up too easily, I don't fall asleep for a long time, or I keep waking up", and, possibly item 323 "I spend much of the day lying down to rest", would result in a dramatic reduction in \hat{p}_z (almost to zero) with only a modest reduction in scale length. However, applying wave 2 results to wave 1 data produces a much less dramatic effect. A small reduction in \hat{p}_z is obtained when items 328 and 327, "I sit around half asleep" are excluded. Perhaps in conclusion, one might only consider discarding item 328 from scale construction. Table 7.13 illustrates the same methodology applied to the "recreation" scale.

TABLE 7.13:

Using individual p values as a criterion for eliminating items from a category scale: recreation

(a) Using wave 1 results on wave 2

Items in (descending) rank order of \hat{p} value	318	319	329	316	322	317	321	315
No of items in scale	8	7	6	5	4	3	2	-
\hat{p} value (x 1000) obtained for wave 2	27	27	28	29	70	57	39	-

(b) Using wave 2 results on wave 1

Items in (descending) rank order of \hat{p} value	317	315	322	319	318	320	321	316
No of items in scale	8	7	6	5	4	3	2	-
\hat{p} value (x 1000) obtained in wave 1	64	65	70	76	68	47	46	-

NOTE: average \hat{p} value 70 (x 1000) range 20-155 41 -24-103

The results suggest there is little to be gained in dropping any items from the scale, in terms of their impact on the category score. It would appear that more empirical evaluation and theoretical appraisal is required in this area.

7.3:

Multivariate assessment of selected item sets

Early assessment of item sensitivity to interviewers in section 7.1 was presented in terms of univariate analyses. Section 7.2 utilizes univariate information to explore ways of reviewing p values for category scores. Category scores themselves are cumulative summary statistics constructed on the assumption of underlying unidimensionality for member items. It would only seem reasonable, therefore, to examine the responses to any category or item set as a multivariate data set. By considering the net interviewer biases or α_i 's estimated under "fixed" effects assumptions for 1-way unbalanced analysis of variance designs for each item in a category a matrix of interviewer effects can be constructed. On the basis of an interviewer (observation) by effect (variate) matrix a multivariate assessment can proceed. In this way it is possible to learn more about the way in which different items may be affected by different interviewers. The methodology was first described by O'Muircheartaigh (1976) and adopted to evaluate the annoyance scale for the ANS in O'Muircheartaigh and Wiggins (1981). Any evidence for multi-dimensionality of interviewer effect is initially obtained by multivariate analysis of variance and, if present, subsequently explored via principal components analysis (PCA) of the "effects" matrix. This procedure will be illustrated for one category from the F.L.P in the following section (7.3.1). In addition, since estimates of effects are obtained for unbalanced data the effect of mis-specifying constraints ($\sum \alpha_i = 0$ as opposed to $\sum n_i \alpha_i = 0$) will be illustrated in section 7.3.2. Another issue of practical consequence, namely "rotation" in PCA of the effects matrix, to identify subsets of items and interviewers who exert any dominant influence on

such items is explored in section 7.3.3. Finally, information gained about the structure of any interviewer effect at an early moment in time compared to subsequent factors analysis assessments will be demonstrated in 7.3.4 using confirmatory factor analysis.

7.3.1:

Multivariate analysis of variance: waves 1 and 2 (Wilks Lambda)

The major enhancement in the appraisal of the F.L.P. categories is that we have replicate multivariate analyses for all of the scales. Table 7.14 summarises the MANOVA results. There is consistent evidence for the presence of multidimensional interview effect in five categories across the waves, namely Body Care, Ambulation, Recreation, Sleep and Rest and Emotion. Eating, Mobility, Household Management and Communication are more indecisive, being "significant" in only one of the two waves. Finally, Work, Alertness and Social Interaction appear reasonably untroubled by the differential impact of interviewers across both waves.

TABLE 7.14:**Multivariate analysis of variance: waves
1 and 2 (Wilks Lambda)**

category	WAVEONE		WAVETWO	
	approx F value	p level	approx. F value	p level
Eating*	1.39	.013	.96	.577
Body Care	1.35	.001	1.281	.007
Ambulation	1.38	.005	1.355	.009
Mobility	1.57	.000	1.087	.272
Work*	.76	.933	1.138	.242
Household Mgmt	1.10	.241	1.207	.090
Recreation	1.90	.000	1.779	.000
Sleep and Rest	1.37	.022	1.623	.001
Communication	1.10	.235	1.318	.030
Alertness	1.10	.235	1.176	.123
Emotion	1.64	.000	1.428	.007
Social interaction	1.05	.316	1.009	.459

* reduced item sets

Sleep and Rest exhibit consistent evidence for multivariate effect across both waves. On grounds of expediency we will pursue the implications of this finding, rather than examine all of the scales in detail. For more appraisal the reader is referred to Wiggins (1985).

It is interesting to examine whether or not different items in the scale are affected in different ways by the interviewers and to what extent such findings are consistent between the waves. That is, are the same groups of items sensitive to the same group of interviewers? Alternatively, if certain interviewers contribute most of the effect for different items are the same interviewers and items involved for both waves?

To examine the structure of interviewer effect a PCA is carried out on the "effects" matrices for each wave respectively. The "12x7" matrix of effects for wave 1 and the "11x7" matrix for wave 2 (interviewer 3 was sick) are presented for reference in table 7.15 below.

TABLE 7.15:
Interviewer effects matrices for sleep and rest items:

(a) wave one

Sleep and Rest interviewer effects matrices under

$$\sum \alpha_i = 0$$

Item/ Interviewer	323	324	325	326	327	328	329
1	.1762	.2500	.0885	.1189	.0008	.1746	.1230
2	-.0113	-.0625	.0010	-.0686	-.0496	-.1496	-.0645
3*	-.0023	.0714	.0814	.0474	.1294	.0468	.0872
4	-.0367	-.1759	-.0615	.0170	-.0492	-.1143	-.0530
5	.0762	.0500	.0385	.1689	.0429	.0746	.0230
6	-.0238	-.0500	-.0115	.0189	.0008	.0246	-.0270
7	-.0738	.0500	-.0615	-.0811	.0008	.2746	.1230
8	-.0738	.0000	-.0615	-.0311	-.0492	.0246	-.0270
9	.0032	-.0192	.0154	-.1311	.0277	-.1485	-.0501
10	.0171	.0227	-.0160	-.0857	-.0037	-.0436	-.0816
11	-.0738	-.1500	-.0615	-.1311	-.0492	-.2254	-.1270
12	.0373	.0278	.0496	.0911	.0619	-.1143	.0396

* Not in Wave two

TABLE 7.15:

**Interviewer effects matrices for
sleep and rest items:**

(b) wave two

Sleep and Rest interviewer effects matrices under

$$\sum n_i \alpha_i = 0$$

item/ interviewer	323	324	325	326	327	328	329
1	-.0281	-.0108	-.0169	.0471	-.0281	-.1854	-.0130
2	-.0281	-.1166	-.0169	-.1067	-.0281	-.0604	-.0899
4	.0164	.0029	.0029	.0054	-.0281	-.0076	-.0232
5	-.0281	.0251	-.0169	.0933	.0386	.0813	.0434
6	.0488	.1430	-.0169	.0471	.2796	.2761	-.0130
7	-.0281	-.0416	-.0169	.0266	-.0281	-.0521	-.0232
8	-.0281	.0441	.0546	-.0353	-.0281	.0289	.0530
9	-.0281	-.1082	-.0169	-.0401	-.0281	-.1854	-.0232
10	.0719	-.1416	-.0169	-.0067	-.0281	.2146	.1101
11	.0830	-.0194	.0943	.2266	-.0281	-.1854	.2434
12	-.0281	.0918	-.0169	-.0591	-.0281	.0527	-.0899

The results of the PCA for wave one (table 7.16 (a)) indicate that two components account for almost all of the variation in the interviewer effects for the seven item scale (76.1%); for wave 2 (table 7.16 (b)) a similar picture emerges with two components accounting for 72.4% of the variation in the interviewer effects. Tables 7.17 (a) and (b) present component scores for each interviewer on the two main components for each wave.

TABLE 7.16:

**Principal component analyses on
matrices of interviewer effect for sleep
and rest category (scale)**

(a) wave one

Component	Eigenvalue	Variance contribution	Cumulation
1	4.02	57.3	57.3
2	1.32	18.8	76.1
3	.96	13.7	89.8
4	.47	6.7	96.5
5	.15	2.2	98.7
6	.06	.9	99.6
7	.02	.4	100.0

Note: Wilks Lambda = 1.37, p<.05

(b) wave two

Component	Eigenvalue	Variance contribution	Cumulation
1	2.92	41.7	41.7
2	2.22	31.7	73.4
3	.98	14.0	87.5
4	.48	6.9	94.4
5	.24	3.5	97.8
6	.14	1.9	99.8
7	.02	.2	100.0

Note: Wilks Lambda = 1.623, p< .01

TABLE 7.17:

**Principal component scores for the
first two components identified in
table 17.16**

(a) wave one

Interviewer	Component	
	I	II
1	2.06	.62
2	- .43	.82
3	1.05	-1.49
4	-1.00	- .03
5	.81	.53
6	- .21	.04
7	.05	1.88
8	- .66	.84
9	- .38	-1.11
10	- .36	- .26
11	-1.61	- .61
12	.70	-1.23

(b) wave two

Interviewer	Component	
	I	II
1	- .29	- .59
2	-1.10	- .68
4	- .20	- .07
5	.04	.46
6	.22	2.67
7	- .42	- .38
9	.18	- .16
9	- .63	- .97
10	.45	.09
11	2.64	- .75
12	- .91	.37

In wave 1 (table 7.17a) most of the variation that the first component describes is attributable to opposite net biases of the pairs of interviewers (1 and 3) and (11 and 4). Whereas in wave 2 (table 7.17b) most of the variation described by the first component is attributable to interviewer 11, more or less standing alone, in opposition to interviewers 12 and 2. In this sense we can say that interviewer variance results mainly from the fact that certain interviewers tend to substantially distort the responses in one direction on average whilst others distort the responses in the opposite direction. Obviously one can continue with this approach for all of the main components identified. The next consideration is to attempt to identify subsets of items which are similarly affected by the interviewers. This has been carried out by calculating the correlation between component scores and the α_i s. It could also be conducted by a direct inspection of the α_i s.

TABLE 7.18:

**Clusters of items identified by PCA of
"effects" matrices on Sleep and Rest
items.**

(a) wave one

Cluster of items correspond to components (range of correlations)

1 (323) I spend much of the day lying down to rest
(324) I sit for much of the day
(325) I sleep or doze most of the time - day and night I
(326) I lie down to rest more often during
the day (.48 to .91)
(327) I sit around half asleep
(329) I sleep or doze more during the day

2 (328) I sleep less at night, for example I wake
up easily, I don't fall asleep for a long
time, or I keep waking up

(b) wave two

Cluster of items correspond to components (range of correlations)

(323) I spend much of the day lying down to rest
(325) I sleep or doze most of the time - day
1 and night I
(326) I lie down to rest more often during (.77 to .96)
the day
(329) I sleep or doze more during the day

2 (324) I sit for much of the day
(327) I sit around half asleep II
(328) I sleep less at night, for example (.72 to .92)
I wake up easily, I don't fall asleep for
along time, or I keep waking up

The two clusters of items exhibited different patterns of variation in the $\hat{\alpha}_i$ s for both waves. With the exception of items 324, 327 and 328 there are a consistent core of four items which make up part/whole of the first cluster across both waves. These four items (323, 325, 326 and 329) are interpretable in the sense that they all are concerned with sleeping more whereas the other three items are concerned with sleeping less or being drowsy. The group of interviewers associated with the first cluster of items across both waves is not homogenous, indicating that these core items were sensitive to the effect of a fairly widespread group of interviewers. Although interviewer 11 appears to play a prominent role across both waves. The second cluster of items identifies item 328 "I sleep less at night ..." as pulling away from the majority of items, standing alone in the first wave but joined by items 324 and 327 in the second wave. Again this item (328) involves a different group of interviewers in both waves; (7,2) versus (3,12 and 9) in wave 1 and (6) versus (11 and 9) in wave 2. In conclusion, then, for this category there appears to be two consistent cores of items (323, 325, 326 and 329) and 328 which largely involve different interviewers within each wave; the interviewer associated with each group of item(s) changes between the waves; with the exception of interviewer 11.

The disparities between the waves are neatly underlined by the following correlations; firstly the correlation across "effects" for each item between waves and secondly, the correlation across "effects" for all interviewers by each item between waves. These correlations are given in table 7.19.

TABLE 7.19:**Correlations of "effects" between waves
for Sleep and Rest items.****(a) Interviewers across items between waves****Interviewer Correlation**

1	-.15
2	.34
4	-.44
5	.30
6	.34
7	-.66
8	.39
	only interviewer to see same respondents
9	.61
10	-.40
11	.37
12	-.55

Note: Interviewer 3 not in wave 2

(b) Items across interviewers between waves**Item Correlation**

323	-.31
324	.00
325	-.55
326	.00
327	.01
328	.01
329	-.51

The illustration provides further evidence of inconsistency over time. Few interviewers produce consistent influence between waves, and this, in turn is reflected in the zero (or near zero) and negative correlations reported for items across time, ie. there is no straightforward replication of cross-sectional analyses at wave 1. Similar conclusions were also reported for Body Care, Ambulation and Emotion (Wiggins 1985) and may well indicate a need for thorough investigation of item content, working, instruction and interviewers' interpretation be conducted prior to any field work. Wiggins (1985) did report that for other categories, e.g Recreation that the same item groups do involve the same interviewers across time and that occasionally the majority of items within a category involve the predominance of a particular interviewer., eg. Ambulation, interviewer 11. These last two observations should lead to a closer inspection of interviewer characteristics and styles. Many interviewers may be conspicuously absent from exerting a strong influence and sound developmental piloting would enable supervisors and researchers to contrast styles and understanding.

All of the multivariate analyses reported so far has been based on the assumption of properly specified constraints ($\sum n_i \alpha_i = 0$) to estimate effects. In the following sub-section the impact of mis-specifying these constraints (ie. using $\sum \alpha_i = 0$ as for balanced data) is considered.

7.3.2:

An illustration of mis-specifying constraints when estimating interviewer effect.

For unbalanced data in a single factor design the appropriate constraint for estimating interviewer net biases is $\sum n_i \alpha_i = 0$ this results in the estimating α_i 's by subtracting the "mean of the interviewer means" ($\sum \bar{y}_i / m$, for m interviewers) from each interviewer mean for any particular response. Under $\sum \alpha_i = 0$, as in the case of balanced data, the effect would be to subtract the overall mean or "grand mean" ($\bar{y}_{..}$) from each interviewer mean. Thus the magnitude of any estimated effect would simply be a reflection of the separation of the "baseline" means, the "grand mean" and the "mean of the interviewer means" for any particular item. As table 7.20 shows below for wave 1 data on the "Sleep and Rest" category, the correlation of the two sets of means is very close ($r = .9994$)

TABLE 7.20:

A comparison of "baseline" means used to estimate interviewer effects : Wave 1 items in Sleep and Rest

Item no.	323	324	325	326	327	328	329
Mean of Interviewer Means	.0750	.2512	.0616	.1256	.0468	.2279	.1242
Grand means	.0738	.2500	.0615	.1311	.0492	.2254	.1270

Not surprisingly for this case there appears to be negligible difference in the impact of mis-specified constraint when replicating a PCA analyses of the "effects" matrix under $\sum n_i \alpha_i = 0$. Two components account for almost all of the variation (again 76% as in table 7.16(a)) and the same clusters of items and reported interviewer influences were observed as in tables 7.17(a) and 7.18(a). Extending the analyses to wave 2 "effects" under $\sum \alpha_i = 0$ again showed little impact. The correlation between "baseline" means was .9952. Continuing this empirical evaluation with data for the "Recreation" category illustrated little to no change for either waves. Again this was largely due to high intercorrelations between "baseline" means across items (.9988 and .9754 respectively). Such findings are comforting, but whilst they do not draw attention to serious mishaps in interpretation they still emphasise the need to carefully examine the nature of what exactly constitutes an estimated effect.

Another aspect of conventional wisdom in PCA analysis is to pause to consider the impact of unrotated versus rotated factor solutions. With small numbers of observations (interviewers) factor/component loadings are likely to be subject to large errors of sampling variability so it may be wise to carefully consider the impact of rotation on interpreting any interviewer influence.

7.3.3:

A review of the practice of rotating axes when determining the pattern of interviewer effects.

Figure 7.5 below plots the principal component scores for each interviewer on the two dimensions (components) identified for wave 1 "effects" for "Sleep and Rest" items. These plots coincide with scores shown in table 7.17(a) for the unrotated solution, they are reproduced below together with scores based on the rotation solution in table 7.21.

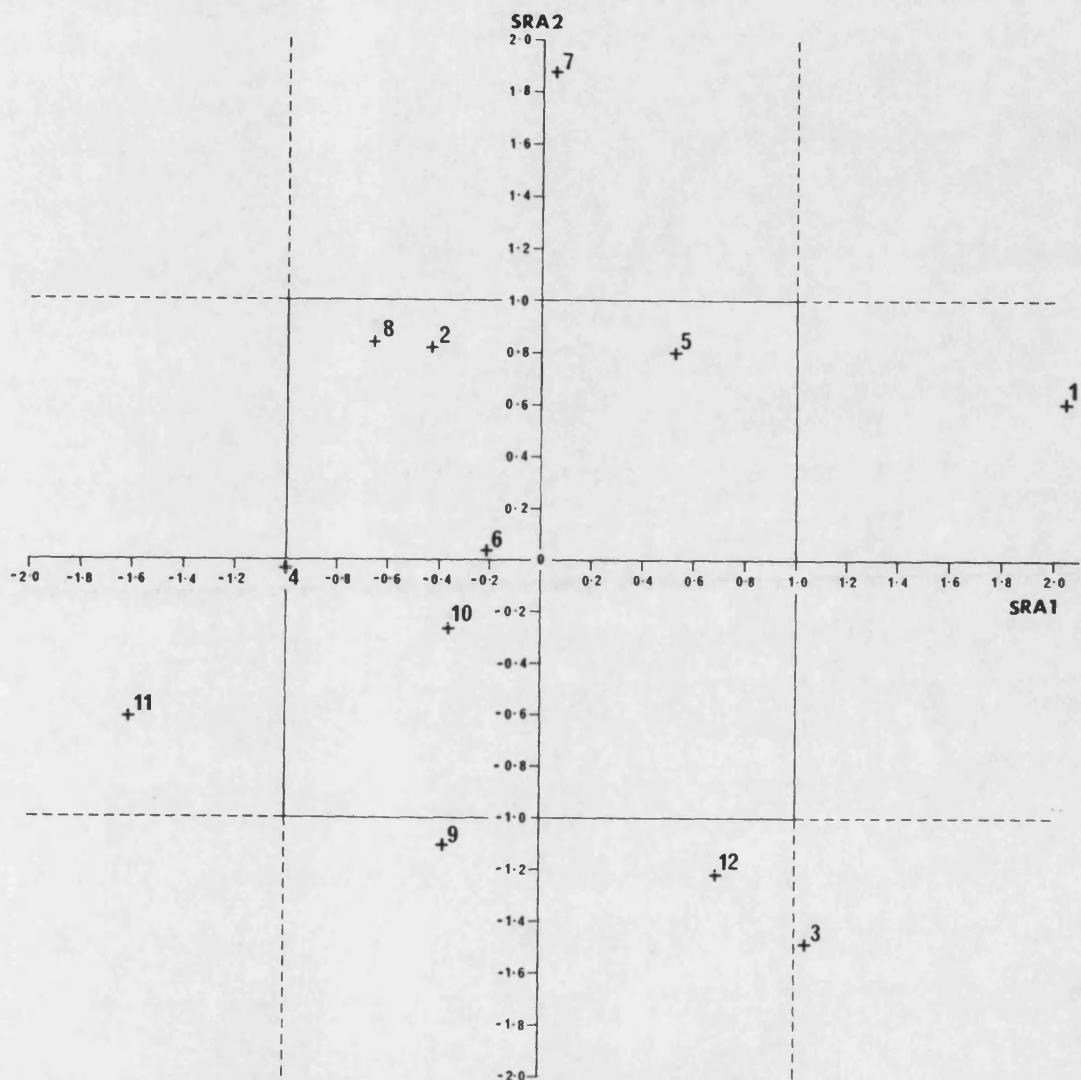
TABLE 7.21:

**Principal component scores for wave 1
"effects'on Sleep and rest : Unrotated and
rotated solutions.**

Interviewer	<u>UNROTATED</u>		<u>ROTATED</u>	
	I	II	I	II
1	2.06	.62	1.16	1.81
2	-.43	.82	-.86	.35
3	1.05	-1.49	1.77	-.45
4	-1.00	-.03	-.73	-.68
5	.81	.53	.27	.93
6	-.21	.04	.19	-.11
7	.05	1.88	-.1.18	1.46
8	-.66	.84	-.1.05	.21
9	-.38	-1.11	.43	-1.09
10	-.36	-.26	-.11	-.44
11	-1.61	-.61	-.83	-1.51
12	.70	-1.23	1.33	-.48

Figure 7.5:

Plot of principal component scores for interviewers in wave 1 for "sleep and rest" item effects: Unrotated solution

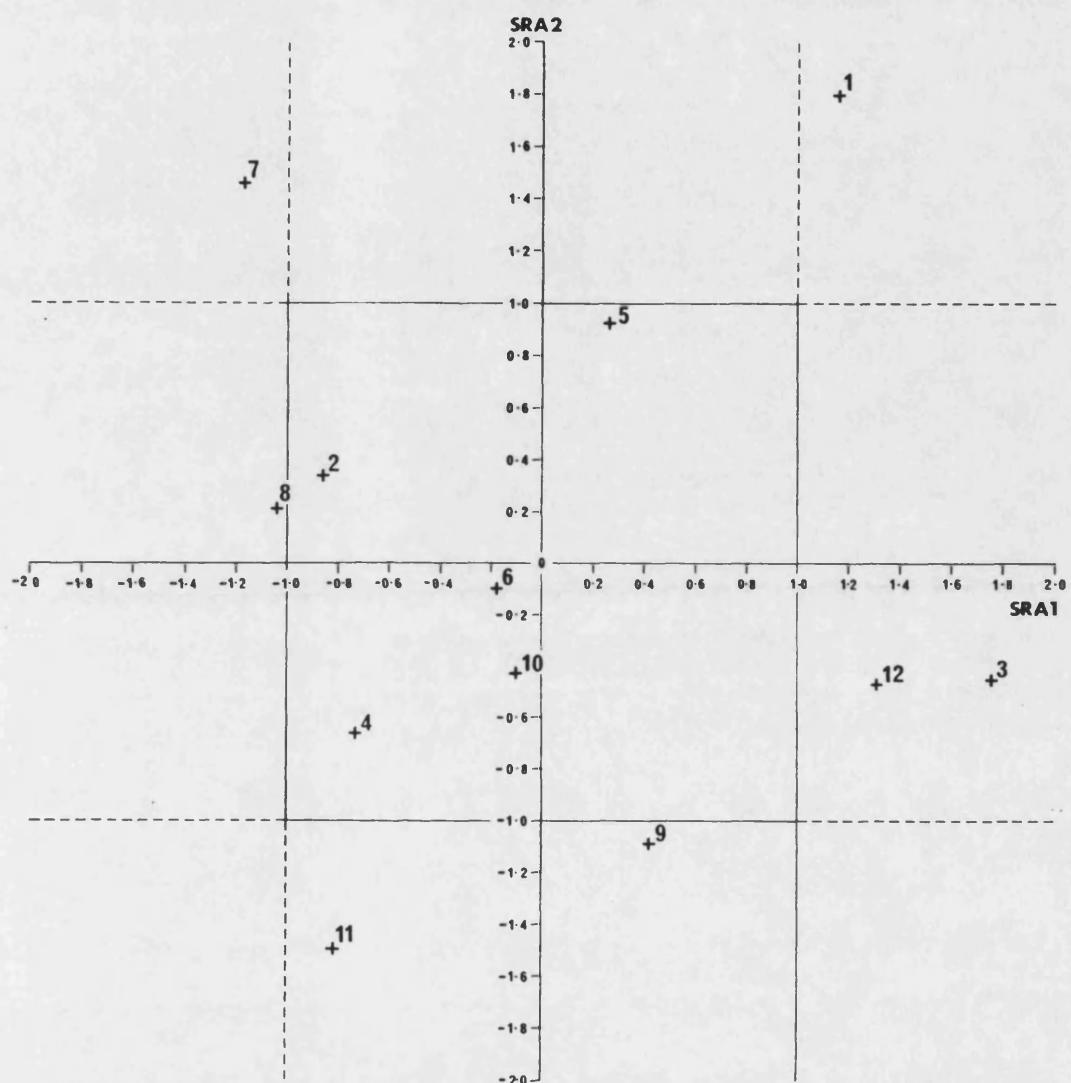


NOTE: PC scores are standardised (0, 1).

Box delimits + or -1 standard deviation along each component.

Figure 7.6:

Influence of rotation. (see transparency inset)



NOTE: PC scores are standardised (0, 1).

Box delimits + or -1 standard deviation
along each component.

In figure 7.5 a "square" has been marked around the origin to help identify any score outside of a range + or - 1 standard deviation (as all scores are standardized). The configuration confirms the earlier appraisal that interviewers 1 and 3 are pulling in an opposite direction to interviewers 4 and 11 on one axis and 7 stands in opposition to 3, 9 and 12 on the other axis. These influences are associated with two clusters of items as reported in table 7.18(a). In order to appreciate the impact of rotation it is useful to use the inset transparency (Fig 7.6). Centering the diagram on the origin of Figure 7.5 enables the viewer to witness the "shift" in relative position for particular interviewers. Notice the influential shifts for interviewers 1,3,7 and 11 and 12 which also results in slightly different subsets of items being identified; notably items 324 and 329 now join 328 as a "second" cluster. Table 7.22 provides the correlations of PC scores with "effects".

TABLE 7.22:

Correlations of principal component scores with effects (under $\sum n_i \alpha_i = 0$) for Wave 1, Sleep and Rest.

Under Rotation

Item	Component		
	I	II	
323	.68	.47	
324	.59	.71	*joins 328 as distinct from unrotated
325	.92	.26	solution
326	.57	.48	
327	.79	.13	
328	.18	.94	
329	.48	.73	* also joins 328, as distinct from unrotated solution

Under rotation it would now appear that interviewers 1,3 and 12 behave differently to 4,7 and 11 for the first subset of items, with 1 and 7 versus 9 and 11 for the second subset. What seems to be borne out by the above comparison is that a group of predominant interviewer effects can be ascertained, namely those of interviewers 1,3,4,7,9,11 and 12, some of whom are more marginal with regard to direction of effect than others. Also items 324 and 329 are more "marginal" than the remainder as to their cluster membership.

These observations are supported by direct inspection of the "effects" matrix. For instance consider items 323, 325 and 326 which were consistently identified as belonging to a subset under both solutions (also having high item effect inter correlation, average .71). Interviewers 1 and 11 are almost always at opposite ends of the spectrum on these items. However, for interviewers 3 and 7 who join them under the rotated solution the separation is not so obvious; table 7.23 reproduces the relevant effects.

TABLE 7.23:
Selected interviewer effects for three
Sleep and Rest items.

Item	323	325	326
Interviewer			
1	.1762	.0885	.1189
3	-.0023	-.0115	.0189
7	-.0738	-.0615	-.0811
11	-.0738	-.0615	-.1311

A similar evaluation for wave 2 data was also conducted. Interestingly, the "rotation" solution produced no material changes to the interpretation, as reported in 7.3.1. However, bearing in mind the caution about sample size and the results for wave 1, it would seem wise to accept that, in general, there will be both interviewer and item marginality. It is recommended that analysis proceed by initial consideration of both rotated and unrotated principal component score plots and final confirmation of any interpretation in the "effects" matrix itself.

Major differences in interviewer influence and item marginality over time for "Sleep and Rest" have already been noted in 7.3.1 however, it would seem appropriate to attempt to discover how well the knowledge about the structure of wave 1 interviewer effect performs in predicting the pattern of wave 2 results. This is now addressed in the final sub-section.

7.3.4:

A strategy for confirmatory factor analysis

Maxwell (1966, chapter 6) suggests that on occasions investigators may wish to specify the factorial composition of a set of variates ie. posulate in advance the number of factors (dimensions) and the pattern of factor loadings. Ideally, a test of goodness of fit, to see if the factors account for observed correlations between variates, would then be desirable. In this manner the structure and loadings obtained for analysis of wave 1 "effects" could be regarded as exploratory and applied to wave 2 "effects" matrices. As one interviewer (no 3) was not in the second wave an exploratory factor analysis was conducted using the wave one correlations based on the eleven interviewers remaining in both waves. A two dimensional structure was identified (principal components solution accounted for 76% of the total variation). The loadings matrix following a varimax rotation is given in table 7.24 overleaf.

TABLE 7.24:

**Loadings matrix for wave one interviewer
effects based on eleven interviewers
in both waves**

Item no	Varimax rotation solution	
	I	II
323	.85	.34
324	.61	.69
325	.93	.24
326	.59	.44
327	.67	-.08
328	-.10	.96
329	.37	.83

LISREL (Jöreskog K G and Sörbom D, 1986) enables the user to test specific hypotheses about the structure of any correlation matrix. In this instance, a two factor structure with exact loadings derived from the table above can be specified in order to try and reproduce the observed correlations amongst the effects for wave two. A chi-square goodness of fit test for this model takes the value 62.2 with 21 degrees of freedom. An extremely poor fit ($p < .00$). A less stringent model specification might be to simply identify which items might be expected to load on particular factors. This can be done simply from the loadings matrix in table 7.24. Table 7.25 reproduces the loadings structure specified in an alternative model specification.

TABEL 7.25:

**Loadings pattern specified in alternative
confirmatory analysis for wave two
interviewer effects correlations**

Item no	Factors	
	I	II
323	1.0	0.0
324	1.0	1.0
325	1.0	0.0
326	1.0	1.0
327	1.0	0.0
328	0.0	1.0
329	0.0	1.0

Note: in the table above a '1.0' indicates that an item is identified with a particular factor and is 'free' to be estimated in the modelling. Item loadings identified by a '0.0' indicate that the item is constrained or 'fixed' to have a zero loading on a factor.

The chi-square goodness of fit for this model takes the value 54.61 with 12 degrees of freedom. Again a very poor fit ($p<0.00$). For further details about goodness of fit criteria and LISREL see Everitt (1984).

These confirmatory factors analyses would, therefore, seem to suggest that there is little information in the observed and latent relationships obtained in the first wave to enhance our understanding of the second wave relationships. The findings underline the observations in section 7.3, in that whilst certain items may be identified as exhibiting a consistent interviewer effect, different interviewers may be exerting different influences during different waves.

7.3.5 :

A summary of findings

Univariate analyses suggest that around a quarter of all items considered have roh values which are significant at the 0.05 level. Typically sampling variance estimates for such items will underestimate the true value by at least a third or even 75% in some cases. Even with large standard errors accompanying the roh-values the implications are serious; beyond the impact on sampling variability one has to consider all of the ramifications for field work practice and questionnaire design. An implicit assumption that "effects" in some sense cancel out for summary scores has no solid empirical support. Some category scores in the F.L.P did have near zero interviewer variance, for others "pruning" the item membership of a category might produce attractive reductions in interviewer variance, but, in general, the idea of a score with near zero interviewer variability remained elusive.

For replicate analyses, the general level of item sensitivity remained at about the same level, however, there the consistency falters. Roh values were typically poor correlates for the same item, indicating substantial item marginality with respect to interviewer effect. There is no assurance that wisdom gained during the first phase of a longitudinal survey will hold for a subsequent phase.

As the majority of single items considered in this evaluation form part of a category or scale, multivariate assessment of effects would always appear appropriate. PCA is an elegant vehicle to identify subgroups of items for which certain interviewers behave similarly or interviewers, who for whatever reason, form "opposing" tendencies for such items. Such information could form a valuable complement to more qualitative appraisals of interviewer style and performance. Again replication of multivariate assessment suggests that few interviewers or items are consistently "sensitive" or "provocative". Where consistencies are identifiable then action needs to be taken to review the item membership of a particular

category or to probe for a deeper understanding of how certain interviewers develop working practices in the field that may be different from their colleagues. In this way interviewer variance studies should not only be part of "good" piloting but regularly integrated as subsample evaluations during main field work.

Interviewer variance studies, as in the ANS and PHS subsamples, typically involve small numbers of interviewers. It would therefore seem wise when using PCA (a) to inspect both rotated and unrotated solutions, (b) confirm that interpretations of interviewer influence accord with the "effects" themselves and (c), of course, ensure that the "effects" are properly estimated.

Finally, the only sound advice to offer the survey practitioner, is that "like sampling error, interviewer variability won't go away!" It is necessary to ascertain its magnitude but not sufficient. Real sources of variability need to be identified. Here a "bridge" is needed between the cognitive aspects of the interviewer process and the statistical warning signals. Accompanying that exploration are useful quantitative attempts to gain more information about the presence of the interviewer and indicators of his/her performance into any investigation of relationships between variables (often category scores, like the FLP or the GHQ). A quantitative strategy for doing so is illustrated in the next chapter.

Chapter 8: Modelling relationships between variables in the presence of interviewer effect

Contents

- 8.1 Analyses of the Aircraft Noise Survey data**
- 8.2 Analyses of the Physically Handicapped Survey data**
 - 8.2.1 Wave one analysis**
 - 8.2.2 Wave two analysis**
 - 8.2.3 Combined level (both waves) analysis**
- 8.3 Summary of main findings**

8.1: Analyses of aircraft noise survey data

The importance of introducing an interviewer factor into modelling relationships between substantive variables was first illustrated by O'Muircheartaigh and Wiggins (1981). Table 8.1 reproduces the analysis of the explanatory power of sensitivity to noise and a measure of psychiatric status (GHQ) in predicting annoyance caused by aircraft noise.

As described earlier in 5.2.2 (b) all three variables were dichotomous. Logistic regression was used to assess any interrelationships between these variables. The results suggested that the conclusions about the relationships between the variables are not substantially affected when interviewer effects are taken into account. The implication of column 2 is that the inclusion of the interviewers as a factor tends to strengthen the evidence about the relationships.

TABLE 8.1:

**Logit analysis of the proportion
highly annoyed by aircraft noise
in terms of sensitivity and
psychiatric status**

Explanatory variables

	(1) Estimate (std. error)	(2) Estimate (std. error)
Psychiatric status	0.376 (0.351)	0.729 (0.390)
Sensitivity	1.095 (0.520)	1.236 (0.553)
Grandmean	0.454 (0.160)	1.81 (0.620)
Interviewer 2	-	-2.021 (0.731)
Interviewer 3	-	-2.376 (0.729)
Interviewer 4	-	-0.569 (0.772)
Interviewer 5	-	-2.305 (0.738)
Interviewer 6	-	-1.479 (0.730)
Interviewer 7	-	0.044 (0.873)
Interviewer 8	-	-1.568 (0.731)
Residual sum of squares	53.39	20.82
Degrees of freedom	25	18

source: O'Muircheartaigh and Wiggins (1981)

Variance components analysis provides an alternative approach to the analysis of these data. The hierarchical nature of data collection is acknowledged by defining nested clusters or 'achieved workloads' of individual respondents (the first level) within interviewers (the second level). In table 8.2 the response variable is binary (0, no annoyance; 1, annoyed) and the quasi-likelihood adaptation is used. The variance for respondents has to be constrained to 1.0, by analogy with the generalized linear models (GLIM). See Longford, 1988b).

Column 1 of table 8.2 presents the results of an analysis including only the fixed effects and ignoring the hierarchy defined by the interviewers (single level analysis).

Column 2 of table 8.2 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: $\hat{p} = 0.408/1.408 = 0.29$. 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.

TABLE 8.2:

**Analyses of aircraft noise survey (ANS)
data: annoyance (0,1) as dependent
variable**

Explanatory variables

	(1) Fixed effects (std. error)	(2) Fixed effects (std. error)	(3) Fixed effects (std. error)
Psychiatric status	.372 (.339)	0.562 (0.346)	0.528 (0.397)
Sensitivity	1.059 (.468)	1.055 (0.485)	1.081 (0.483)
Grand mean	-0.147 (.065)	0.108 (0.264)	0.110 (0.255)
Random effects source	Variance Sigma (std. error)	Variance Sigma (std. error)	Variance Sigma (std. error)
Respondent	1.000 1.000	1.000 1.000	1.000 1.000
Interviewer	-	0.408 0.639 (0.218)	0.371 0.609 (0.211)
Slope for psy	-	-	0.292 0.540 (0.535)
Slope for sens	-	-	0.003 0.053 (0.540)

NOTE: In column 3 parameter estimates for "psy" and "sens" refer to the difference between the second level and the first level for each factor.

This variance components model corresponds to a logit analysis of covariance with interviewers as the classifying factor. Introducing the interaction between this factor and an explanatory variable has a direct analogue as 'variance' 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 3 shows the effect of having 'psychiatric status' and 'sensitivity' in the random part of the model.

The results indicate that those who are GHQ positive are more likely to be annoyed by aircraft noise than those who are GHQ negative, even when their self-assessment of sensitivity to aircraft noise is taken into account. The relative odds calculated from column 2 suggests that for the high sensitivity group the probability of being annoyed is about 11% higher for the GHQ positive respondents (expected value of proportion annoyed ($p = 0.848$ vs. 0.767)). 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 (0.54) 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 itself is an estimate with estimated standard error of 0.535. The value of sigma for sensitivity assessment (.053) indicates reasonable stability of the main effect; though again it carries an estimated standard error of 0.54. Thus, a suitable model for prediction was considered to be one that included psychiatric status in the random part.

The five interviewer level variables described in section 5.2.2(b) were entered into this model one at a time. Their relative impact was judged in terms of the relative magnitude of their effect. The sex of the interviewer and the average number of calls made per workload appeared to have negligible impact on the response, age of interviewer was moderately interesting though not convincing enough to be included. Response rate and years of experience did appear more convincing and are included in table 8.3 overleaf.

TABLE 8.3

**The effect of including interviewer
characteristics in the fixed part of the model**

Explanatory variables	(1) Fixed Effect (std. error)	(2) Fixed Effect (std. error)
Psychiatric status	.473 (.382)	0.586 (0.423)
Sensitivity	1.059 (.478)	1.086 (0.496)
Interviewer's response rate	- -	-4.691 (2.928)
Interviewer's experience	- -	-0.197 (0.089)
Grand mean	.079 (.255)	0.071 (0.205)
Random effects source	Variance Sigma (std.error)	Variance Sigma (std. error)
Respondent	1.000 1.000	1.000 1.000
Interviewer	.373 .611 (.181)	0.180 .424 (0.169)
Slope for psy	.260 .510 (.405)	0.379 0.613 (0.373)

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 'interviewer-level' 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 8.3 contrasts the analyses with or without interviewers' characteristics in the fixed part of the model.

In table 8.3, column 1 gives the results with the interviewers' characteristics excluded and column 2 shows the effects of including interviewers' response rates and experience in the model. Three 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 2 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 (variance components ratio - $0.180/1.180 = 0.15$). In other words a substantial proportion of the interviewer variance can be explained in terms of achieved response rate and years of experience. Third, the effect of psychiatric status as measured in the fixed part of the model (coefficient = 0.586) is still subject to considerable variation between interviewers. The value of sigma is 0.613 with an estimated standard error of 0.373.

Two conclusions emerge from table 8.3. First, from a methodological standpoint, it is useful to discover that measurable interviewer characteristics such as achieved response rate and years of experience can be introduced directly into the model and that within the framework of the overall model they account for a substantial proportion of the variance due to interviewers. Second, despite this, the residual variation due to interviewers suggests that the findings for the effect of psychiatric status is unstable across interviewers.

8.2 The analysis of the physically handicapped survey

The context for analysing the PHS data was described in section 5.3.2 (b). The results for analysis of the modelling the relationships for each wave separately are presented before the "combined level" analysis. In a structural sense, then, variance component models for waves one and two are analogous to the one described in the previous section, where interviewers define the second level of the hierarchy. The major difference being that the F.L.P is regarded as a continuous response variable with a normal error structure.

8.2.1 Wave one analysis

Table 8.4 (column 1) gives the results of fitting the model

$$y_i = \sum_k \beta_k x_{ki} + \epsilon_i \quad (8.1)$$

to the data from first year of the PHS. This model contains only the fixed part referred to in section 3.9. There is a strong effect for work status and self-assessment of health; slightly weaker evidence of an effect for attendance at hospital outpatient services and little evidence of either an age or sex effect.

In column 2 the analysis is extended to include the effect of the interviewers. The model becomes

$$y_{ij} = \sum_k \beta_k x_{kij} + v_j + \epsilon_{ij} \quad (8.2)$$

where $j = 1, \dots, 7$ represents the interviewers. This is analogous to the analysis of covariance with no interactions. 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 and this is explored further below.

Similarly we can consider inclusion of the variable 'age' in the random part of the model.

$$y_{ij} = \sum_k \beta_k x_{kj} + v_{0j} + v_{2j} x_{2ij} + \epsilon_{ij} \quad (8.3)$$

The results are given in column 3. This appears to lead to an improvement of the model, and the consequences are rather interesting. The fixed effects estimate is 0.065 but the square root of the corresponding variance (sigma) is much larger (0.139). Thus for a typical interviewer the slope on age is positive but the variation of this slope is so large in comparison that there are 'many' interviewers whose slope on age is negative. Figure 8.1 illustrates the residual effects due to age; it shows that minimal variation (between interviewers) occurs for respondents around the age of 40 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 function occurs for age of about 40 years.)

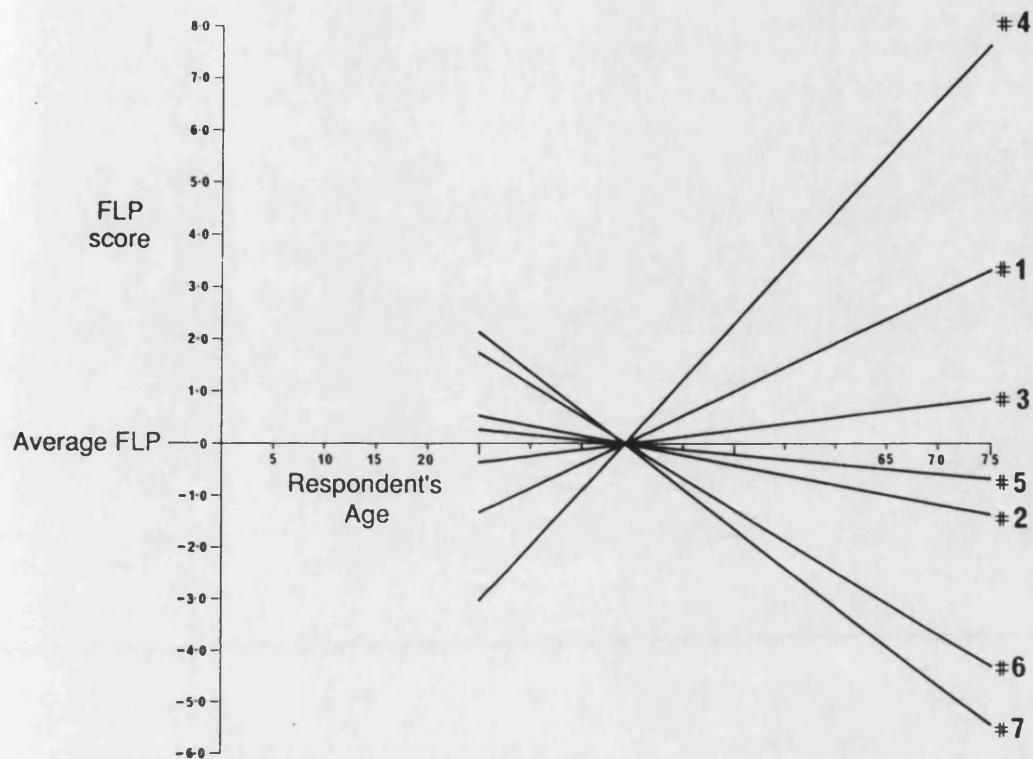
In terms of the impact of individual interviewers we see that the greater part of the variation arises from the contrast between interviewers 1 and 4 on the one hand and 6 and 7 on the other. The results suggest that if interviewers 1 and 4 were to carry out all the interviews, a strong positive fixed effect for age would appear, whereas if interviewers 6 and 7 only were used, there would be an apparent negative fixed effect for age.

TABLE 8.4

Analysis of data from physically handicapped survey: functional limitation profile (FLP) score as dependent variable for wave one

Explanatory Variables (std. error)	(1) Fixed Effect (std. error)	(2) Fixed Effect (std. error)	(3) Fixed Effect
Sex	1.10 (1.96)	1.04 (1.92)	1.32 (1.86)
Age	0.048 (0.078)	0.050 (0.077)	.065 (.090)
Work	4.91 (2.24)	5.04 (2.21)	4.64 (2.14)
Ass. Health	6.30 (1.25)	6.15 (1.22)	6.24 (1.20)
Hosp. Out	-5.38 (3.27)	-5.78 (3.19)	-5.78 (3.07)
Grand mean	-8.32 (0.954)	-7.59 (1.22)	-8.37 (1.22)
Random effects source	Variance Sigma (std. error)	Variance Sigma (std. error)	Variance Sigma (std. error)
Respondent	76.4771 8.74	66.51 8.48	66.51 8.16
Interviewer intercept	-	4.33 2.08 (1.35)	4.86 2.21 (1.09)
Slope for age	-	-	.019 .139 (.074)
Deviance		601.27	597.24

Figure 8.1: An illustration of allowing age of respondent to enter the random part of the model



Note: for the analyses presented in chapter 8 interviewers have to be numbered consecutively. Original numbers used in chapter 5 to describe the experiment are in the left hand column. SCPR field work numbers are in brackets. The conversion is as follows:

4	becomes	6	(9600)
5	"	1	(8090)
6	"	2	(9240)
7	"	3	(9260)
11	"	7	(9905)
12	"	5	(9560)
8	"	4	(9390)

As in the previous section, interviewer level variables were introduced into the model in table 8.5 (col. 2). An interviewers' age, sex or experience rating failed to provoke any pronounced impact ¹. However, average call back means and attitude towards the disabled looked interesting. Table 8.5 provides a complete summary.

The magnitude of the standard error for the effect of call back means undermines any major generalisation - in that the effect on the FLP of increasing calls to obtain a response appears to vary considerably between interviewers. An interviewers attitude score appears more consistent, in that a movement of around 4 points on the ATDP scale would correspond to a 1 point change on the FLP scale (ATDP scores ranged from 66 to 89, high scores denoting tolerance).

¹. Footnote: details of computer runs not included here.

TABLE 8.5:

**Two level analysis for phs wave one with
two interviewer level variables in the fixed
part of the model and with age respondent as
a random effect**

Explanatory Variables	Fixed effect (standard error)
Ave calls	.24 (2.83)
Att. dis	.25 (.18)
Sex	.94 (1.88)
Age	.04 (.10)
Work	4.23 (2.15)
Ass. health	6.12 (1.21)
Hsp. Out	-5.72 (3.08)
Gm	-26.99 (1.07)
Random effects source	Variance Sigma (std. error)
Respondent	66.93 8.18
Interviewer	2.32 1.52 (.98)
Slope for age	.03 .16 (.07)
Deviance	593.94

NOTE: Slope by intercept covariances have been omitted.

8.2.2: Wave two analyses

Table 8.6 (col 2) contrasts the "fixed effects" model for year two with the corresponding year one (col 1), previously presented in table 8.4 where interviewers define the second level in the hierarchy.

TABLE 8.6:

2 level analysis of PHS data for separate waves. FLP as the dependent variable

Explanatory Variables	WaveOne (1) Fixed effect (std. error)	Wave Two (2) Fixed effect (std. error)
Sex	1.04 (1.92)	.42 (1.96)
Age	0.050 (0.077)	.12 (.07)
Work	5.04 (2.21)	5.75 (8.14)
Ass. health	6.15 (1.22)	6.49 (.99)
Hosp. Out	-5.78 (3.19)	-3.73 (5.16)
Grandmean	-7.59 (1.22)	-15.22 (.96)
Random effects source	Variance Sigma (std. error)	Variance Sigma (std. error)
Respondent	71.90 8.48	76.70 8.76
Interviewer intercept	4.33 2.08 (1.35)	0.00 0.00 (2.48)
Deviance	601.27	602.93

Similar main effects are present as for wave one, with "age" demonstrating a slightly stronger presence. Surprisingly, perhaps, there appears to be no variation to account for among the interviewers themselves.

Apart from noting that the standard error for sigma is around 2.5 there may well be a case for abandoning further pursuit of the influence of interviewer characteristics. However, in the interests of symmetry this was carried out, producing some findings worthy of note, though full details of all of the runs are not provided. The average number of calls per interviewer told a similar tale to wave one, with some variation of effect among interviewers ($\sigma = 0.14990$, fixed effect -0.25042). In contrast to wave one attitudes towards the disabled appeared almost negligible in wave two (fixed effect $= -0.09525$, $\sigma = 0.07973$). This may well reflect a tendency for the first year's interviewing experience to narrow the range of attitude scores (21 points compared to 33), with 5 out of the 7 interviewers dropping in their total score. Also in contrast to wave one the age of the interviewer now appeared to introduce some effect (fixed effect $= -0.16186$, $\sigma = 0.03187$). There was weak evidence for the existence of a "sex difference" among interviewers but with only two male interviewers this was taken lightly; there also appeared to be fairly wide variation in scores according to experience ratings.

For the combined/longitudinal analysis of the PHS data 84 respondents remained in both waves for seven interviewers as detailed in table 5.7. Two modelling approaches were considered for the hierarchy defined in Fig 6.1. For the first approach interviewer characteristics are examined by including year by 'characteristic' interactions in order to examine the relative stability of their possible impact across each wave. In the second approach the effect of individual interviewers over time is examined by including an interviewer factor and a 'year by interviewer' interaction term in the model. Clearly interviewer characteristics will confound the nature of an 'interviewer' effect. However it was not possible to separate out all of these terms in the estimation due to aliasing of terms. The next subsection presents the results for these two approaches. The actual parameter estimates are given in tables 8.7 and 8.8 respectively.

8.2.3: Combined level (both waves) analysis

The outcome of the first strategy outlined in the previous subsection is presented in table 8.7 below.

TABLE 8.7:
Combined level analysis of PHS data; FLP as a dependent variable

Explanatory variables	Fixed effects parameters	
	Estimate	(standard error)
Year	73.3379	36.4510
Ave call	3.0155	2.0453
Att dis	0.5399	0.2030
Age	-0.5132	0.2603
Sex	4.6058	3.5096
Experience	8.2197	3.7699
Sex	0.6130	2.0122
Age	0.1017	0.0750
Work	1.7465	1.2521
Ass. health	3.0209	0.6536
Hsp. Outpatient	-3.4323	1.6263
yr*attitude	-0.7892	0.3474
yr*age	0.9644	0.4602
yr*sex	-2.1263	4.7958
yr*experience	-12.4282	6.3801
gm	-68.0090	0.9751
Random effects source	Variance	Sigma (std. error)
Measurement	16.5283	4.0655
Respondent	71.8217	8.4747 (0.7328)
Deviance	1138.7940	

Principal main effects noted for explanatory variables in both waves are consistent, with a notable "year" effect. Given the way that "hierarchy" has been defined the large respondent level variability is comforting (c.80%), measurement variation can be interpreted in terms of a pronounced year effect, and inconsistencies brought about by the passage of time, interviewer effect or a combination of all three possibilities. The "years by interviewer characteristics" interactions appear to reflect some of the wave disparities noted earlier in 8.2.1 and 8.2.2.

As an alternative strategy interviewer presence can be declared by means of a factor; interaction with "year" of interview also allows inspection of how interviewer effect may differ for different sets of respondents. From table 8.8 we see that these year * interviewer interactions are much larger than the individual interviewer estimates! Only interviewers 4 and 5 have their effects in both years of the same sign relative to interviewer 1 (e.g. for interviewer 2 we have $-0.7050 + 10.1080 = 9.4030$ in year 2 compared to -0.7050 in year 1 and so on). Relative to other interviewers, interviewers 6 and 7 produce strong differential effects in both years, with interviewer 2 showing a large effect in year 2 only. It must be noted that there is no practical reason that for all fixed effects estimates for categorical variables to be made in relation to the first level. However, the observations made here still serve to demonstrate great variation in the direction of effects between the waves.

TABLE 8.8:

**2 level analysis for combined waves
declaring interviewer as nested factor**

Explanatory variables	fixed effects parameters		
	Estimate	St. Error	
Year	-4.2638	3.4394	
Interviewer			
1	0.0000	0.0000	
2	-0.7050	3.6576	
3	-1.6658	4.0436	
4	1.0275	3.5378	
5	-2.5006	4.0206	
6	-4.7037	3.4157	
7	-8.3664	3.2807	
Sex	0.9638	1.9100	
Age	0.0987	0.0722	
Work	2.0077	1.2315	
Ass. health	2.9871	0.6399	
Hsp. Outpatient	-3.6841	1.5900	
yr * Interviewer			
1	0.0000	0.0000	
2	10.1080	5.3515	
3	2.2422	5.4027	
4	-0.1375	3.7936	
5	1.3693	5.2428	
6	7.0002	6.4904	
7	12.3625	5.1301	
gm	0.4208	0.9230	
Random effects source	Variance	Sigma	St. Error
Measurement	15.8976	3.9872	-
Respondent intercept	63.7586	7.9849	0.6970
Deviance	1126.2433		

8.3: Summary of main findings

The analyses presented in this chapter illustrate the potential of variance component analysis for incorporating interviewer effects in data analysis and for developing an understanding of the nature of interviewer variability. The modelling strategies enable estimation of 'typical' main effects whilst at the same time allowing consideration of these effects to vary among different interviewers. This is a distinct advantage over simply including interviewers as a 'factor' in the modelling. Additionally, interviewer characteristics can enter the model to potentially explain any interviewer variance.

The first application (section 8.1) produces results analogous to analysis of variance and confirms the presence of a substantial interviewer effect. The value of the interviewer intraclass correlation coefficient of 0.29 suggests that the potential impact on the analysis of relationships could be overwhelming. However comparisons of the fixed part of the model with and without the interviewer factor show that in this case the interviewer variability simply masks the strength of the relationships. Introducing 'interviewer-level' variables demonstrates that about 50% of the variability associated with the interviewers can be explained in terms of their individual experience and response rate.

The second application replicates variance component modelling for each wave for the FLP score in terms of a number of explanatory variables. The first wave analysis confirms the presence of substantial interviewer effect; but this does not seriously distort the interpretation of the fixed part of the model. There is further evidence to suggest interviewer characteristics, notably the average number of call backs and their own attitude towards the disabled, help explains around 40% of the underlying variability. There is a further indication that the impact of the interviewers on the responses varies systematically with age. For wave two, different influences emerge. At first sight it might appear that interviewers have little to no differential impact on the responses. However a large standard error for the variance component estimate prompted further investigation. As a result, average number of calls appeared interesting, as did an interviewer's age, though for this wave an interviewer's attitude towards the disabled did not.

Analysis of the PHS data for the combined years (section 8.2.3) generally confirmed the separate wave analyses. Namely, the nature of interviewer effect generally varies between years and the impact of interviewer characteristics is different between the years. The exception here being 'average number of calls' that an interviewer makes. The apparent stability of this finding is interesting, but tempered by a large standard error in both waves.

Part One: Conclusion

The measurement of social phenomena means assigning numbers to some population or sample of elements, in accordance with a set of rules. Using a structured questionnaire to achieve this by means of an interview requires skill both in terms of the design of the questionnaire and its implementation. Interviews will differ in skill and respondents will differ in motivation. The social context in which the interaction (or interview) is played out plays a crucial role in determining the quality of the survey data. Typically, researchers will attempt to carefully reduce any potential distortion in the response patterns by piloting, attention to question wording, the questionnaire length, its' saliency and use of language as well as recall aids and briefings. But, when it actually comes to 'information getting' it is the interviewer her/himself who brings meaning to the experience. Each person in the interview may have fixed attitudes and/or stereotypes of others. Both respondent and interviewer have visible characteristics which may create a false sense of security or hostility. The extent to which these influences translate themselves into quantifiable distortions in the response patterns predominates the concern of this part of the thesis. We have seen that it is not simply enough to assume that any biasing effects will cancel out. In order to demonstrate these findings it is necessary to establish a mathematical framework to explore the nature of response errors. This in turn, requires modifications in the survey design (Mahalonobis (1946), Hansen, Hurwitz and Madow (1953), Deming (1960), Kish (1962). If the presence of 'interviewer effects' can be properly demonstrated then the work of social psychologists, notably Cannell (1954 onwards) and Sudman and Bradburn (1976) can usefully inform researchers as to how to study the sources of such variation (Chapter 2). It would be unusual to embark on research with the sole purpose of studying interviewer variability. The necessary modifications to a survey design can be achieved to different degrees of symmetry and completeness (Kish 1962). Where studies have been mounted they typically employ one way classification or nested designs. Factorial designs are less common. Table 3 (chapter 2) conveys a range of possible design modifications for a number of major studies of interviewer variability in terms of parameters used by Hansen, Hurwitz and Madow (1953) to illustrate the potential for the study of interviewer effect. In principle, increasing use of centralised or regionalised telephone interviewing field forces

make studies of interviewer variability readily attainable. Design sophistication need not deter researchers. In fact, the regular use of 'interpenetrating samples' for both face to face and telephone interviewing presents itself as an attractive compromise between methodological interests and budgetary strictures.

Kish (1962) made important contributions to the study of interview variability. He drew attention to the use of traditional analysis of variance techniques to estimate components of interviewer variability for one way classification schemes where interviewers completed an unequal number of interviews. Relatively small interviewer variance components can lead to dramatic increases in variance for modest sized workloads. The use of variance component estimates of interviewer variability to obtain 'roh', the intra class correlation coefficient have endured in univariate assessments of interviewer effect (Collins, 1978, 1979, 1980).

For the ANS, results indicate that for many of the attitudinal items the variance of the estimates derived from the survey will be inflated by a factor of between 1.6 and 3.6 due to the presence of an interviewer effect. Moreover, the two scales involved, the GHQ total score and the annoyance score, also show evidence of dramatic effects. For the 135 FLP items evaluated in the PHS study (Wiggins, 1985) in both years around 15% of the items lead to an inflation of the variance estimates between 1.6 and 5.2. Jack-knife methodology (Wolter, 1985) is illustrated in chapter 7 to provide appropriate variance estimates for roh.

Although average values of roh are similar in both waves of the PHS study (.03, .025 respectively) many individual items exhibit considerable variability between the years, i.e. the same item will not always be sensitive to interviewer effect. Also, of interest, is the relationship between individual items and the summary scale scores those groups of individual items generate. O'Muircheartaigh (1976) first explored this relationship by examining the relationship between the magnitude of roh associated with the scale score and the simple average of the individual roh values for items constituting the summary score. This approach was replicated for each of the twelve FLP scale scores in both waves of the PHS study. About half of the scale scores had reasonably low roh-z values despite the fact that individual roh values within the scales were high. The evidence is less conclusive for other scales; though 'sleep and rest' and 'recreation' appeared consistently

sensitive to interviewer effect. By adapting a methodology first applied by McKennell (1977) attempts were also made to look for ways of reducing the number of items in a scale in terms of the magnitude of their individual *roh* values and the resulting impact on the *roh-z* for any modified summary score based on fewer items. Near zero values of *roh-z* were obtained for 'sleep and rest' by dropping two items, but no reduction was possible for the 'recreation' score. The approach may well be usefully applied during the developmental phase of questionnaire design. The realisation of the concept of a summative scale score rests on the assumption of 'unidimensionality', that all of the items in scale measure the same underlying phenomena. This assumption can normally be evaluated by treating the items as a multivariate item set. It is therefore logically consistent to analyse interviewer variability in this context. Individual item by interviewer effect matrices can be constructed for any set of items using an appropriate 'analysis of variance' model to estimate effects. These 'effects' matrices are then subjected to a principal components analyses, following evaluation strategies demonstrated in O'Muircheartaigh (1976) and O'Muircheartaigh and Wiggins (1981). These strategies enable the 'dimensionality' of any interviewer effect to be established; for instance, to what extent do interviewers behave uniformly? Any departures from unidimensionality enable different subsets of items with closely correlated interview effects to be identified. By inspecting principal component scores it may also be possible to identify subgroups of interviewers who have differential impact on different items. The procedure is illustrated for two scale scores, 'sleep and rest' and 'recreation' in chapter 7. There is strong evidence that similar groups of items may be sensitive to interviewer effect over time, but often those same clusters are generated by different individual interviewers; occasionally the same item group is identified with the same interviewer over time (Wiggins, 1985). These observations indicate that even with small groups of interviewers, the sets of items which show similar effects form interesting and interpretable clusters. The implications for training and supervision are enormous. More information on interviewer characteristics, attitudes, personality and interviewing style may help to begin to determine why it is that some interviewers are conspicuously absent from provoking any noticeable effect and others are not. Video feedback and interviewing the interviewers might also unravel why some items are consistently sensitive to an effect, and why others only occasionally.

The (experimental) design context necessary for the exploration of interviewer variability has important consequences for the analyst. The founding principles of analysis of variance need to be addressed in the light of the interview context. Estimates of variance components and fixed effects, will be based on unequal achieved interviewer workloads. The assumption that all analyses are based on unbalanced designs has direct consequences for estimation and hypothesis testing.

Chapter 3 presents the reader with a review of these implications (largely based on Searle, 1971, 1987) so as to create a 'guide' to matters that are of immediate concern to the researcher. Time did not permit all of these implications to be demonstrated in chapters 7 and 8. However, it is hoped that the potential investigator will be better informed about what is feasible. In this way, researchers may gain more from their experimental investment. The appraisal of interviewer variability should not begin and end with an inspection of F-ratios. Ultimately, researchers will have to work harder to provide useful hypotheses and fuller information about possible sources of variation. Questionnaire items are typically categorical; polytomized responses are at best 'ordinal', often reduced to binary outcomes for analysis. Estimates of ρ_{oh} assume that response measures are quantitative. There is clearly a tension between conventional assessment of item sensitivity to interviewer effect and the level of measurement used to record a response. For binary items where the proportion endorsing a particular category varies between 0.2 and 0.8 there is no major cause for concern (Cox, 1970). As a caution against simply using point estimates or ρ_{oh} for binary responses the proportion endorsing the item has been routinely provided alongside estimates of sampling variability (see tables 7.8 and 7.9). Where proportions fall outside the recommended range there is little to do other than abandon serious appraisal. In principle, summative scores based on multi item sets, e.g. the GHQ and FLP, appear to be more convincing candidates for treatment as quantitative measures. The GHQ score can be thought of as a discrete or 'count' variable, and the FLP as a percentage score based on the proportion of the maximum 'weighted' score attainable. However, even where the researcher can feel reasonably happy with the measurement level assumption, other fundamental assumptions underpin traditional analysis of variance. Namely, equality of variance, normality and independence of the observations. All too often these issues are paid scant attention;

simple checks advocated by Scheffe (1959) may lead to proper remedial action, e.g. the use of transformations.

Polytomized items, like individual ratings for separate GHQ items and the annoyance scale suggest that alternatives to traditional analysis of variance might be appropriate, e.g. the use of contingency coefficients (Gale and Kendall, 1957 and Everitt, 1977) or log linear analysis (Payne, 1977). These approaches are briefly introduced in chapter 7. Generally, there appears to be a temptation for researchers to 'plough on' with analysis of variance, rather than to carefully consider ways in which appraisal could match the measurement context. One potential application in the review of multivariate item scales might be to devise 'distance' measures to summarise the degree of closeness between items for each pair of items within a scale (e.g. see Charlton, 1981). Separate 'distance' matrices could then be constructed for each interviewer as well as the who sample so as to facilitate 'three way multidimensional scaling' (or individual differences scaling, Krzanowski, 1987). This would enable the investigator to identify which interviewers produce a similar effect for particular items as well as how particular interviewers stand in relation to one another. The method has the attraction of not ignoring the measurement level assumption. It would also 'sidestep' the need to consider how to estimate an interviewer effect as in O'Muircheartaigh (1976) and chapter 7. This latter consideration brings our attention directly on to the substance of chapter 3, namely how to 'contextualise' our analysis of variance model and decide whether or not the interviewer effect conveyed in any model is 'random' or 'fixed'?

Under a completely randomised one way classification is it always appropriate to assume that interviewers are represented by 'random' effects as though they were selected from a large pool of potentially available participants? In large fieldwork organisations supervisors may often persuade or select groups of interviewers to participate in an 'experiment' on the basis of their individual experience or commitment to the organisation rather than on sampling considerations. It is the authors view that it would be pragmatic to regard interviewer effects as 'fixed', unless there are convincing grounds to do otherwise. If the experiment were repeated the same interviewers would be included. Under such conditions any conclusions drawn only apply to the interviewers included in the experiment. Conventionally, conclusions are made on the basis of the magnitude of 'mean'

square errors'. Under one way classifications the expected values of mean square errors for the fixed and random effects models are similar. This is not the case for more complex designs. Chapter 3 reviews the lessons from the theory of the linear model under fixed effects assumptions for three basic experimental designs : one and two way classification schemes as well as nested (or interpenetrating designs) under the assumption of 'unbalancedness'. Under the one way classification individual terms in the model are not by themselves estimable unless 'restrictions' or 'constraints' are included in the model specification. Usual constraints or 'sum effects zero' ($\sum \alpha_i = 0$) have different implications for estimates we than will, 'weighted sum effects zero'. Constraints are included as a means of obtaining a solution for estimating individual terms, whereas restrictions are said to be an integral part of the model. In practice, this distinction appears so subtle as to be almost elusive. Certainly, it is not always easy to argue the case the inclusion of a set of constraints on 'intuitive' grounds as well as on their 'algebraic' appeal. The use of two different sets of constraints ($\sum \alpha_i = 0$ versus $\sum n_i \alpha_i = 0$) was applied in the context of multivariate assessment of the 'sleep and rest' scale in section 7.3 for the construction of item by effects matrices. In that instance little noticeable difference resulted in the interpretation of the findings. There is no reason why this should always be the case; different choices about the use of restricted versus unrestricted models, restrictions versus constraints will potentially have an impact on the interpretation. For unbalanced one way classifications weighting an interviewer's effect by his/her completed workload size does seem reasonable ($\sum n_i \alpha_i = 0$). The similarity between the findings between effects estimated under this constraint and those estimated under 'sum effects zero' is accounted for by the closeness between the overall mean of the responses and the mean of the interviewer means used under the two separate procedures.

Similar cautions can be made about nested designs: it is impossible to estimate individual terms or attach any meaning to F-ratios without imposing constraints on effects within levels. For example consider a design where interviews are nested within geographical areas i , interviewer effects are described in the model by β_{ij} , where constraints are generally of the type

$$\sum_{i=1}^{b_i} w_{ij} \beta_{ij} = 0, \text{ where } w_{ij} \text{ is typically } \frac{n_{ij}}{n_i} \text{ or } \frac{1}{b_i}$$

For two way classifications the equality of interviewer effects can be tested but the necessary restrictions are 'data dependent' and have no obvious intuitive appeal. The situation could also be complicated by the presence of 'empty cells', e.g. an interview failing to complete an assignment in a particular location. Where all of the cells are filled restrictions of the type $v_{ij} = 0$ enable the equality of interviewer effects to be tested. However 'sum of interaction effects zero' does not have any particular natural appeal. The general problem arising from the resulting analysis of variance for two way classifications with unbalanced data is the nature of the conclusions that can be drawn from the significance or otherwise of F-ratios of the type $F(\alpha/\mu)$, $F(\beta/\mu, \alpha)$, $F(\beta/\mu)$ and $F(\alpha/\mu, \beta)$ implicit in tables 7. $F(\beta/\mu)$ and $F(\alpha/\mu, \beta)$ are not used for the same purpose; $F(\alpha/\mu, \beta)$ tests for the effectiveness of adding α -effects over and above μ , whereas $F(\alpha/\mu, \beta)$ tests the effectiveness of adding α -effects in it already. The tests are not the same and cannot be 'loosely' referred to as 'testing α -interviewer effects'. Provided simultaneous fitting of both α and β has an explanatory value for the variation in the response Searle (1971) shows that the process of determining whether, in fact both α and β are needed in the model requires consideration of at least 16 possible outcomes (see table 3.5). Needless, to say these concerns do not arise with balanced designs.

Added complexity for interpretation and estimation of any interviewer effect with increasing design sophistication seem to make the suggestion for the use of interpenetrating designs both appealing on grounds of theoretical simplicity as well as practical consideration. These designs could be accompanied by careful use of planned comparisons or contrasts between interviewers. These contrasts could be used to test for the influence of certain background characteristics, e.g. gender, race or status.

Another, seldom tried alternative to the use of complex designs might be the use of cell-means models (or μ_{ij} models). These strategies overcome the problems of interpretation encountered by imposing 'data dependent' restrictions or constraints with elegant algebraic simplicity. 's' sample means resulting from the combination of each level of any classifying factors (interviewer and/or location) are used as unbiased estimators of population means from which the original observations are deemed to be a sample. There is no confusion as to what functions are estimable,

that their b.l.u.e's happen to be and exactly what hypotheses are being tested. Models are always of full rank. The outstanding substantive problem is specify hypotheses of interest in terms of the population means themselves. From the early literature (e.g. Speed, 1969) it would appear that the temptation might be to relate hypotheses about functions to conventional analysis of variance formulations rather than functions and resulting restrictions that actually relate to the context of the data (Searle, correspondence, 1986).

Cell mean models apart model description underlying the analysis of variance need not be exclusively presented as all effects 'fixed' versus all effects 'random'. Even the simplest one way classification could be considered as a combination of fixed (μ) and random (α) effects, and thus described as a 'mixed' model. Whenever we consider designs where at least one of the effects is considered to be random our attention will focus on the estimation of variance components as well as fixed effects. Defining interviewer effects as 'random' in any model formulation would be directly analogous to using ρ_{oh} as the proportion of total variance attributable to the interviewers themselves (Kish, 1962). The necessary ingredients are variance components. Estimating these components for unbalanced data relies on several possible procedures. We have already seen for a two way classification under 'fixed' effects assumptions that there are two analyses of variance possible (see table 3.5). Thus there is no uniquely obvious set of sums of squares or quadratic forms that can be used for variance estimation. With balanced data the method of estimating variance components rests on equating expected values of mean squares to their observed values in the analysis of variance for the corresponding fixed effects model (labelled by Searle as 'the analysis of variance method').

Chapter 3 presents a review of other variance component estimation methods for unbalanced data (namely, Henderson's method 2, fitting constants, the analysis of means and the symmetric sums method). All of these methods simplify to the analysis of variance method when the data is balanced, but there is no obvious criteria to help choose an appropriate method when the data is unbalanced. By considering all models as 'mixed' effects models, i.e. having at least one effect fixed (μ), Searle (1971) develops a strong argument for the development of a unified estimation procedure to allow the simultaneous estimation of both variance components and fixed effects. The most attractive option

was 'maximum likelihood estimation'. Unity did not however spell 'simplicity'. The algebraic complexity identified with traditional analysis of variance for unbalanced data was not jettisoned. Explicit maximum likelihood estimators were unobtainable. Hartley and Rao (1967) developed a set of equations from which specific estimates are obtained from iteration, and Searle (1970) derived large sample variances for MLE variance components despite the absence of explicit estimators. Dempster et al., (1977) formalised the EM algorithm as a general procedure for MLE in a wide variety of models which include variance component models. A modification of the ML approach leads to REML (restricted maximum likelihood, Patterson and Thompson, 1971) to ensure unbiased estimation of variance components. Aitken et al., (1981) used GENSTAT for both ML and REML estimation; both methods led to similar conclusions. Anderson and Aitken (1985) developed an important application to interviewer variability for binary responses in a variance component model for an unbalanced interpenetrating design (random allocation of respondents to interviewers within geographical areas). Computational difficulties associated with the EM algorithm led Goldstein (1987) and Longford (1986d) to separately develop algorithms which do not require the inversion of large matrices and converge rapidly. Longford (1987) presents details of a Fisher scoring algorithm for unbalanced nested random effects models. The resulting software package (VARCL) is reviewed in chapter 3 and demonstrated for analyses presented in chapter 8 for situations involving both quantitative and binary responses. Data has to be hierarchical or nested. For one way classifications respondents are nested within interviewers. Interviewers define the second level in the nesting hierarchy. For an interpenetrating design where interviewers were arranged within geographical areas, area would define the third level in the hierarchy. Variation in the response variable attributable to the differential impact of the interviewers can be estimated by a variance components. The possible influence of other substantive respondent variables could also be included in the model. Their influence on the response as explanatory variables would be estimated as 'fixed' or 'typical' effects alongside the variance component estimates. These typical effects can also be allowed to vary according to which interviewer carried out the interview i.e. they enter the random part of the model. The other exciting aspect of this approach is that specific characteristics at each level of the hierarchy can be measured and included in the model. In this manner, variables like and individual interviewer's response rate can be included in the model. Their combined impact

can be evaluated in terms of the resulting reduction in the interviewer variance component, i.e. as possible explanations or sources of interviewer variability. Thus traditional analysis of variance techniques become embedded in the VARCL framework.

The first application of VARCL in chapter 8 uses the quasi-likelihood adaptation of the estimation procedure in VARCL for a binary response variable. Essentially the analysis is an extension of the results presented in O'Muircheartaigh and Wiggins (1981) for the ANS study. The basic analysis confirms the presence of a substantial interviewer effect; around 29% of the variability could be attributed to the influence of the interviewers themselves. This would seem to suggest that the potential impact on the analysis of relationships could be overwhelming. However, a comparison of the fixed part of the model with and without the interviewer factor show that interviewer variability simply masks the strength of the relationship. A similar interpretation was provided by O'Muircheartaigh and Wiggins (1981) using GLIM. Where VARCL demonstrates its' potential is by showing how variance component models can be used to identify sources of interviewer variability, in particular, variation in interviewer experience and interviewer response rate can be seen to account for about 50% of the variability introduced into the responses by the interviewers.

In the second application, the method is demonstrated for quantitative response variables, namely the FLP score in the analysis of relationships for a set of explanatory variables used in Charlton (1981). The analysis was replicated separately for each wave of the PHS study. In the first wave, again substantial interviewer effects are demonstrated, but the simple interviewer effect does not contaminate the interpretation of the fixed part of the model. Variation in the average number of calls completed by each interviewer and their own attitude towards the disabled (using the ATDP scale) did appear to account for about 40% of the variability introduced into the responses by the interviewers. There is a further indication that the impact of the interviewers on the responses varies systematically with the age of the respondent. For the second wave it did appear at first glance that there was no evidence for the presence of any interviewer effect. In spite of a zero variance component estimate, its' large associated standard error estimate suggested some further investigation of the influence of interviewer characteristics would be advisable. There seemed to be further evidence to consider the influence of the average number of call completed by

each interviewer but, other influences looked unconvincing.

The final application involved the adaptation of the definition of hierarchy to handle repeat measures for the PHS study. With one exception each interviewer involved in the experiment received a different random allocation of respondents in the second wave (see chapter 6). This meant that extending the hierarchy to include a third level, 'time', was not feasible. To some extent the problem was circumvented by defining respondents as the second level in the hierarchy each with two nested responses, one for each wave. In this manner, interviewer effect is included in the fixed part of the model as a factor. The results generally confirming the findings observed for the separate waves. The nature of the interviewer effects varied between the years. Two interviewers, 4 and 5, were the exception; they exerted similar distorting influences in both years. Otherwise, the influence of any interviewer tended to reverse in direction between the years. The only consistent source of interviewer variability appeared to be the average number of completed calls; the size of its' accompanying standard error tempers any sweeping generalisation. Further, it was interesting that interviewer's own attitude towards the disabled ceased to play any major influence after one year's interviewing experience of the PHS survey. Indeed, for the second wave all scores tend to move downwards and closer together to indicate 'less tolerance' on the attitude scale. Variance component models provide a valuable 'new' approach to the analysis of the impact of interviewer variability. VARCL establishes a framework which unites the traditional analysis of variance with exploration of possible sources of interviewer variability by allowing the investigator to include 'second level' variables (interviewer characteristics) as part of the model. Allowing variables to enter the random part of the model also provides an indication of the sensitivity of particular items (fixed effects) to different subsets of interviewers. Thus the concerns of the data analyst and the survey methodologist can be incorporated under a single approach.

PART TWO

**The impact of non response bias and a global evaluation of
the extent of its' effect**

Introduction:

Procedures for estimating the impact of non-response bias go some way towards treating the overall design of a survey as a single unit. Provided proper attention is given to the estimation of the effect of components of variance due to sampling and non-sampling error in a survey design a global term can be constructed to reflect all sources of survey error. It is appropriate to refer to this criterion of accuracy as "mean square error" (Cochran, chapt.1 1953).

Typically levels of accuracy will depend on the costs of obtaining information. A decision to increase or improve levels of accuracy will increase costs. Survey design criteria must therefore recognise this interdependence. "Efficient" survey design is a "trade-off" between costs and desired levels of accuracy, or maximum levels of accuracy for a fixed budget. It is therefore desirable to combine cost and mean square error in a single criterion to reflect the quality of information in a survey item.

The great variety of survey expenditure represents a complex reality, and, detailed information is seldom available in survey reports. Kish's observation from 1965 is still pertinent today, "ordinarily the sampler has no precise data on cost factors, and must base his decisions on estimates or guesses a good cost model helps to ask the right questions and to make good guesses .. ". The illustration that follows in Chapter 13 is no exception, costing detail was not made available to the researcher. This makes a thorough review of what exactly constitutes a "good cost model" essential.

The data serving the illustration (the Occupational Mobility Survey) was attractive because it provided sound information on the difficulty of obtaining a response by recording call-back attempts and whether or not the interviewer arranged appointments with potential respondents. Although common fieldwork practice, such information is rarely recorded or made available. Conveniently in this instance the availability of such information facilitated a retrospective evaluation of (approximate) total survey error which concurred with modelling strategies and software availability for estimating non-response bias. It is recognised, however, that there are competing methodologies for the estimation of non-response bias. A number of these are considered in the review chapter immediately following this introduction. What this section of the thesis does is provide a framework for evaluating survey designs. Until more imputation techniques are used as common practice generalizable findings will not be possible.

In summary, chapter 9 reviews methodologies for estimating non-response bias and chapter 10 explores various approaches to the definition of an appropriate costing model. Chapters 11 and 12 anchor the specific survey for illustration, and chapter 13 provides the case study to demonstrate the particular methodologies reviewed in chapters 9 and 10.

chapter 9: A review of the effect of non response and strategies for exploring its' impact on survey estimates

9.1 Effects of nonresponse

9.2 Types of nonresponse and some empirical results

9.2.1 Types of nonresponse

9.2.2 Trends in nonresponse

9.2.3 Factors affecting nonresponse and characteristics of nonrespondent-interviewer interaction

9.3 Data collection remedies for lessening impact

9.3.1 An overview

9.3.2 Call-backs

9.4 Data processing techniques for lessening the effect of nonresponse:an operational review of some approaches

9.4.1 Adjustments for bias without repeated call-backs: the indirect approach

9.4.2 Post hoc imputation procedures which take no account of call back policy: the passive approach

9.4.3 Adjustments which take account of call-back policy: the direct/analytical approach

9:1 Effects of nonresponse

The term nonresponse when applied to a potential respondent refers to the failure to obtain any information for item responses. Such non-productive effort is referred to as "unit nonresponse" as distinct from "item-nonresponse" or "partial nonresponse". These latter are terms used to denote failure to obtain some of the characteristics of a particular respondent. Partial nonresponse will not be considered in detail, although it does provide the motive force for the development of data imputation techniques. Some of these methodologies will be reviewed in section 9.4.2.

Nonresponse is present in both censuses and surveys. For convenience consider a population divided into two distinct strata. The first consists of all respondents who furnish measurement and the second of potential respondents for which no measurement is obtained. Let N_R and N_{NR} be the numbers of individuals in the two strata and let $W_R = N_R/N$ and $W_{NR} = N_{NR} / N$, be the respective proportions of respondents and nonrespondents in the population strata so defined. If \bar{Y} represents the population mean, the amount of bias in the respondent mean \bar{Y}_R is

$$\bar{Y}_R - \bar{Y} = W_{NR} (\bar{Y}_R - \bar{Y}_{NR}) \quad (9.1)$$

Thus the bias is a product of two components: the difference between the respective means in the strata and the proportion of nonresponse. Clearly fieldwork strategies to minimise W_{NR} are well justified otherwise we have to place bounds on the value of the respondent mean from sources external to the investigation. Cochran (1953) suggests that with a continuous variable the only bounds that can be assigned with certainty are "often so wide as to be useless". Consider the case of proportions; the nonrespondent value must lie between 0 and 1 so either the investigator can make extreme assumptions, i.e. set the nonrespondent proportion to 0 or 1 or develop some procedure for constructing confidence limits for P . The table below is reproduced from Cochran (1963) shows how for a series of values of W and P for sample size 1000 a rapid increase in the width of the interval with an increasing proportion of nonrespondents soon develops.

TABLE 9.1:**95% confidence limits for P (%) when N = 1000**

%	nonresponse, 100W ₂	sample percentage		
		5	10	20
0		(3.6, 6.4)	(8.1, 11.9)	(17.5, 22.5)
5		(3.4, 11.1)	(7.6, 16.3)	(16.5, 26.5)
10		(3.2, 15.8)	(7.2, 20.8)	(15.6, 30.4)
15		(3.0, 20.5)	(6.8, 25.2)	(14.7, 34.3)
20		(2.8, 25.2)	(6.3, 29.7)	(13.7, 38.3)
				(37.2, 62.8)

Taken from Cochran (1963)

Birnbaum and Sirken (1950) developed an interesting method of determining a suitable achieved sample size by assuming known values of W_2 from previous surveys and stated values for a tolerable level of desired error in the sample proportions. The practical implications of their work suggest that any sizable proportion of nonresponse make it difficult to guarantee high levels of precision by increasing the sample size of the respondents (refer table 13.3 Cochran). Their work confirms the importance of allocating a substantial proportion of effort to minimising W_{NR} .

Typical survey reports provide no information about the nonrespondent stratum. As response rates are reported and survey estimates are based solely on respondent values it is implicitly assumed that the nonresponse bias is "zero", i.e. $\bar{Y}_R = \bar{Y}_{NR}$. Where information does exist for nonrespondent characteristics there is a strong suggestion that such an assertion is unwise. First consider two examples from mail surveys with personal interview follow-up of nonrespondents after the third mailing. The first example is taken from Cochran (1963) and the second, from Locker et al., (1981).

TABLE 9.2: Responses to three requests in a mailed inquiry

	No. of Growers	% of Popln.	Ave. Number of Fruit Trees per Grower
Response to first mailing	300	10	456
Response to second mailing	545	17	382
Response to third mailing	434	14	340
Nonrespondents after 3 mailings	1839	59	290
total population	3116	100	329

From Cochran (1963)

TABLE 9.3: Per cent disabled in each stage of survey (individuals aged 16 and over)

	Disabled	Non-disabled	Total
1st mailing	16.6 (1309)	83.4 (6588)	100.0 (7897)
2nd mailing	15.8 (595)	84.2 (3176)	100.0 (3771)
3rd mailing	11.5 (436)	88.5 (3363)	100.0 (3799)
Student follow-up	11.5 (343)	88.5 (2645)	100.0 (2988)
TOTAL	14.6 (2683)	85.4 (15772)	100.0 (18455)

Excludes 289 individuals for whom stage of return not known

From D Locker et al., 1981

In the Cochran example on fruit growing the number of fruit trees per grower declines in successive mailings. In the second example based on a postal survey of Lambeth households, we witness a decline in the number of disabled screened at successive attempts to secure a response. Next, consider the following example based on Thomsen and Siring's excellent review of the "cause and effects of nonresponse". Here a sample of 5047 women aged 18 to 44 years was selected for a major study of fertility in 1977. Data collection was organised in two phases, the first phase consisted of interviewers making up to 8 calls on potential respondents, thereafter trained interviewers were detailed to obtain information for outstanding contacts. To evaluate the effects of nonresponse information regarding the number of live births and income for nonrespondents was found from external sources. Tables 9.4 and 9.5 overleaf summarise the findings:

TABLE 9.4: Fertility among respondents and nonrespondents by age

	age				all ages
	18-24	25-34	35-44		
Mean number of live births among respondents	.40	1.68	2.55	1.57	
Mean number of live births among nonrespondents	.26	1.21	2.10	1.19	

From Thomsen and Siring (1983)

TABLE 9.5: Mean income ¹ among respondents and nonrespondents by age

	age				all ages
	18-24	25-34	35-44		
Mean income among respondents	340.4	757.8	888.8	677.3	
Mean income among nonrespondents	272.4	612.0	862.2	581.7	

From Thomsen and Siring (1983)

¹ N.kr. 100

Fertility is higher among respondents than nonrespondents in all age groups. This is possibly due to the fact that women with children are more available for interview than women without children. Also, for some reason the nonrespondents income tends to be lower than that for respondents among all age groups. The other striking evidence provided is that for both survey variables "refusals" among the nonrespondents tend to be more like the respondent population than other nonresponses. This point is considered in a wider context in section 9.2.3. The main observation here is that the composition of the "two" strata of responses appears to depend largely on the method of data collection and perceived saliency of the subject matter.

9.2: Types of nonresponse and some empirical results

9.2.1: Types of nonresponse

In chapter 1 we referred to the survey population as a collection of individuals that can be surveyed under a specific set of survey conditions. Utilising a scheme proposed by Murthy (1983) it is possible to distinguish between the survey population and the frame population in the following manner:

- i) nontraceable individuals
- ii) missed individuals
- iii) individuals who are temporarily absent
- iv) not at homes for all attempts
- v) refusals
- vi) incapacity to respond
- vii) other nonresponse, e.g. unwillingness to cooperate, lost schedules
- viii) duplication, individuals external to target population

Authors differ with respect to the degree of detail they attach to a classification of nonresponse (e.g. contrast Hansen, Hurwitz and Madow, 1953; Cochran, 1963, Kish, 1965). The differences appear to arise from practical emphasis rather than serious disagreement about the nature of the problem. A fundamental distinction between two principal sources of nonresponse: refusals and not at homes runs throughout the literature. Not at homes arise through a variation in the amount of time spent at home and are felt to be ultimately attainable, indeed controllable with proper planning and call-backs. Refusals are thought to vary in their firmness of rejection. Recent Family Expenditure data from Norway suggests they typically account for around 50% of nonrespondents; GHS data from Britain indicate refusal rates around 12-14% [or about 77% of nonresponse]. Tables from Thomsen and Siring (1983) and Lievesley (1986) underline these observations.

TABLE 9.6: General household survey annual response rate

year	non-contacts %	refusals %	base (total effective sample)
1971	2.7	11.9	15432
1972	2.6	13.5	15307
1973	2.9	13.5	15360
1974	2.3	11.6	14232
1975	2.2	12.0	15327
1976	2.3	11.2	15310
1977	2.1	12.0	15315
1978	2.5	12.8	13957
1979	2.7	11.8	13437
1980	2.4	13.5	13943
1981	2.2	11.6	13939
1982	2.2	11.7	11970
1983	2.6	12.5	11862
1984	3.7	13.7	11867

(Taken from Lievesley, 1986)

TABLE 9.7: Nonresponse rates by components in family-expenditure surveys since 1967

year	size of sample (households)	nonresponse rates (%)	refusals	not at home	other	total (%)
1967	5008	21.8	-	-	-	100
1973	4707	28.6	39	-	-	100
1974	1388	32.6	46	19	35	100
1975	1648	32.3	42	22	36	100
1976	1707	31.0	44	25	31	100
1977	1419	30.0	44	27	29	100

(Taken from Thomsen & Siring (1983))

Deming (1953) distinguishes between permanent and temporary refusals with a view of nonresponses as arising from graded classes members of the universe to be sampled, where classes range from an impregnable core of no possible response up to a class of complete response. The notion of "hard core" nonresponse exerts a major influence on adjustment methodologies for nonresponse bias, whilst the distinction between temporary and permanent nonresponse has strongly influenced fieldwork practice. Often interviewers are instructed to note reasons for refusal to help fieldwork managers make a decision as to whether or not there is a chance that with a reissue to another interviewer a respondent may be persuaded to change their mind. In Britain, SCPR (Lievesley, 1986), conducted an experiment in which interviewers were asked to make a subjective assessment for each refuser as to how likely they thought it is that with a different interviewer calling back a response could be obtained. A random half sample of initial refusals was reissued to second interviewers who had no knowledge of the assessment made by the initial interviewer. Interviews were obtained with 34% of the reissues, but this success rate varied from 72% (for those considered very likely to respond) to 23% (for those considered very unlikely to respond). The overall success rate did not differ materially from the rate typically expected based on a selective reissue policy (around 38%) but did provide useful evidence to suggest blanket or automatic reissue would not be particularly cost-effective.

Other strategies might involve the use of telephone recall, letter writing or reissue callbacks conducted by specially trained interviewers. (Techniques of randomized response procedures, reviewed by Kim and Fluek (1976), for handling items that might provoke embarrassment may be worth considering.

In reports on nonresponse organizations typically report non-contact rates as in table 9.7 above as well as the conventional split between refusers and not-at-homes. Essentially non-contacts will include categories (i) to (iii) and (vii) above as well as not at homes, Lievesley (1986) refers to these other categories as "fieldwork shortfalls" that may arise from operational difficulties, time or cost constraints, or the inability or unwillingness of interviewers to track down potential respondents. Equally category

(vi) above may also be included as a refusal, here Lievesley suggests the practical distinction between unwillingness to respond and incapacity to respond may be blurred. An individual may be too drunk on one occasion to participate but sober and cooperative on another occasion. Readers should appreciate these disparities when assessing recent trends and results in the following subsection.

9.2.2: Trends in nonresponse rates

Systematic appraisal of response rates is perhaps most straightforward for continuous or repeated surveys. Steeh's (1981) examination of rates for the U.S National Election Studies, 1952-76 and U.S. Consumer surveys from 1954-76 provide a valuable foundation. She demonstrates for both assessments clear increases in refusal rates by geographical area and increases in other sources of nonresponse for large urban areas. Lievesley (1986) also provides illuminating data for Britain on the continuous GHS survey, the repeated Labour Force Survey (both government sponsored) and the commercially sponsored continuous National Readership survey. Whilst no obvious trend is apparent, refusal rates have tended to rise in the recent past for the government surveys (c. 13% for 1983/4 cp. 12% for earlier years). Refusal rates for the readership surveys are stable at around 12% for years up to 1983 but show a worrying increase to around 14/15% for 1984/5.

Nonresponse rates by area for Britain's family expenditure survey shows some improvement for the years 1969-74 (see Figure 1, Lievesley, 1986) but particularly stubborn rates c.37% for the Greater London area. Problems with family expenditure surveys are not only specific to Britain, Thomsen and Siring (1983) show powerful evidence of increase in overall rates for the period 1967-1977 in Norway (see Table 9.7 in section 9.2.1).

Data for ad hoc surveys is clearly more difficult to assess as variations may be due to survey-specific factors such as sponsorship rather than changes in the propensity of the general population to respond. Again Lievesley (1986) provides an interesting account of recent SCPR surveys (1978-85) in the table overleaf:

**TABLE 9.8: Social and Community Planning Research
ad-hoc survey response rates**

year	non-contacts %	refusals %
1978	11.0	7.1
1979	10.4	7.6
1980	10.9	4.1
1981	11.6	8.4
1982	10.4	7.9
1983	11.8	8.2
1984	11.9	8.9
1985	9.8	7.1

Based on 94 surveys conducted by SCPR
(Taken from Lievesley, 1986)

There appears an optimistic recession in increasing non-contact and refusal rates during 1985. This may well be due to increased emphasis on the problem of declining nonresponse rates in training and briefing sessions and fieldwork management practice, e.g. interviewers at SCPR are asked not to return documents relating to noncontacts when the minimum of four calls have been made but to retain them until the end of the field work period, recalling whenever they are nearby. Bonus schemes related to high response achievement have also been attempted. Obviously clear information regards measures to control nonresponse are concealed when we simply review global rates of response. Interpretation is assisted if we introduce an element of structure into any appraisal. Thomsen and Siring (1983) divide variables involved in the cause and effect of nonresponse into three categories:

- i) the dependent variable: level of nonresponse
- ii) indirectly controllable variables: supposed to have an effect on the level of nonresponse but only under indirect control of the survey investigator
- iii) controllable variables: variables where the survey investigator has some direct control e.g. selection and training of interviewers, mode of data collection

There are important conceptual differences between variables designated group (ii) or (iii) status though practical difficulties arise in making such a distinction. The table below from Thomsen and Siring attempts to provide such a framework. This frame will serve to guide our appraisal of recent evidence regarding nonresponse effect and nonrespondent-interviewer interaction in the following subsection.

TABLE 9.9: Framework for the analysis of nonresponse data

CONTROLLABLE VARIABLES	INDIRECTLY CONTROLLABLE VARIABLES	DEPENDENT VARIABLES
Selection and training of interviewers	Qualifications of interviewers	Total nonresponse rates
General working conditions for interviewers and terms of employment	Motivation of interviewers	Refusals
Use of introduction letters	Availability of respondents	Temporarily absent
Use of Incentives	Motivation of respondents	Not-at-homes
Use of proxy interviews		other nonresponse
Public relations		
General instructions		
Respondent burden		
Number of visits per respondent		

(Taken from Thomsen & Siring (1983))

9.2.3: Factors affecting nonresponse and characteristics of nonrespondent - interviewer interaction

In a multiple regression analysis of response rates on workload size, years of experience, sex, age and errors occurring during the introduction of a survey for individual interviewers Thomsen and Siring highlight the most significant variable as workload size. A small negative correlation between size of assignment and response rate was observed. This finding confirmed similar analyses by Platek et al., (1977). Work by Thomsen and Siring on "respondent burden" experimented with different duration times for maintaining records of expenditure in the Norwegian Family Expenditure survey (2,3 and 4 week subsamples) provided little evidence to suggest response rate is affected by the amount of work required to complete the task. Other factors identified the use of specifically trained interviewers to reduce initial refusal rates, the attitude of the public towards the usefulness of surveys and the perceived ability of a survey organization to keep information confidential.

Lievesley (1986) provides systematic information on the nonrespondent-interviewer characteristics to help the investigator pinpoint remedial action or consider adjustment strategies. Primarily evidence suggests some contradictions about the idea of a permanent refusal or hard core nonrespondent. Refusals appear situational, the result of inopportune call times rather than any deepseated antipathy towards surveys. Sandstrom (1977) suggests there is no reason not to believe that one and the same person may react differently on different occasions. Van Westerhoven (1978) concludes "the refusers group is not perpetually non-compliant but a group primarily composed of people who sometimes refuse". Unfortunately, as Denise Lievesley indicates a lot of literature that reports on nonrespondent characteristics fails to relate the analysis to specific survey variables (both questionnaire items and organizational variables) which restricts current appraisals to a limited set of demographic characteristics appertaining to nonrespondents. For example, she demonstrates evidence based on 1981 Census checks on Family Expenditure and General household data a tendency for non-response rates to increase with increasing household size. This may go some way to explain under-representation of younger age groups (20-24) and the

elderly (often single person households) in surveys. She speculates the former is due to the persistence of not at homes and the latter high refusal rates. Studies generally fail to make the distinction between noncontacts and refusals in accord with nonresponse characteristics. Patterns are noted such as the tendency for slight improvements in response rates for the higher educated or social groups (the self employed notwithstanding) and the low ranking of London and the South East in regional assessment. Much more systematic information is clearly desirable if the impact of nonresponse bias is to be appreciated. Work on adjustment procedures which focus only on the noncontacts or not at homes at the expense of the refusals could seriously undermine efforts to reduce nonresponse, especially in situations where noncontact biases are expected to offset refusal bias. For example, Thomsen and Siring (1983) provide evidence to suggest that refusals are more like the general population than non-respondents as a whole. Therefore it cannot be safely assumed that biases are compensating.

By standardising refusal rates by interviewer across a number of surveys conducted by SCPR Lievesley illuminates several of interesting factors: interviewers who have been in the employ of the agency for longer than five years tend to have lower refusal rates than those with less agency experience, interviewers of middle years do better than older or younger colleagues, supervisors and interviewers with extra training do better than those with less training or responsibility. Using tape recorders to document doorstep introductions Morton-Williams (1986) suggests that interviewers with high refusal rates could all be criticised for failing to listen and react to the need's of potential respondents., e.g. interviewers who make spontaneous offers to recall at more convenient times obtain higher response rates.

Recent work by Lievesley on non-contact rates is also equally fascinating. It would appear that maximum availability for an interviewer is not necessarily the best indicator of low non-contact rates. Interviewers with restricted availability appear to make better use of their time and call potential respondents during evenings and weekends. Also interviewers who said they were willing to stay away from home, irrespective of whether this was actually necessary or not, tended to obtain lower non-contact rates.

An important consequence of these findings is the need for more research on calling strategies. Models for nonresponse adjustment based on repeated attempts to obtain information are reviewed in Section 9.4, whereas further consideration of data collection remedies is provided in the next section.

9.3: Data collection remedies for lessening impact

9.3.1: An overview

The simplest approach to collecting data from reluctant or mobile respondents might be to "throw money at the problem", (Proctor, 1977). A cash payment to the respondent may be worthy of serious consideration, even a sliding scale of payment based on a scale of reluctance! Apart from such fancy the investigator has recourse to improvements in field procedures. Following Kish (1965) and Proctor (1977) consider the following aspects of fieldwork and design practice.

- i) propriety:** guarantees of anonymity, confidentiality
sponsorship, media publicity
advance notice, letters of introduction
sensitive doorstep introduction,
use of randomized response techniques
for embarrassing questions
- ii) interviewer selection:** experienced or not,
local or not,
matching interviewer characteristics
with respondents or not
- iii) method of data collection:** mail, face to face interview, telephone,
computer assisted or a combination of some
or all of these
- iv) task structure:** degree of structure and guidance for
questionnaire, length of questionnaire
and style of wording.

Proctor (1977) suggests that these procedures could be implemented with varying degrees of four possible delivery styles:

- i) repeated call backs (mailings or dailings)
- ii) subsampling non-respondents; see Kish section 13.5c and application by Hansen and Hurwitz (1946) to following up non-responders of mail inquiries with personal interviews.
- iii) a graduated series of approaches until respondent simply cannot fail to respond! (the "escalation" approach).
- iv) establish some combination of approaches whereby an optimum level of effort is determined which balances non-response control relative to other survey expenses.

Style (ii) has been given a Bayesian formulation, Ericson (1967), and tends to be undermined by appreciable numbers of hard core non-responders. Style (iii) may be aggravated by measurement bias in that successful initial calls differ in field approach to subsequent successful follow up calls. Further consideration is therefore only afforded to a strategic assessment of styles (i) and (iv).

9.3.2: Call-backs

In interview surveys it is typical to specify a number of call backs or a minimum number that must be made on a potential respondent before placing him/her in the non-response stratum. Repeated call backs represent the most successful means of reducing the extent of non-response, especially for the "not at homes". Repeated mailings or dialings also fit into this category. In survey practice the amount of effort required to obtain a satisfactory response tends to be a function of intuition and past experience rather than considered appraisal. For instance the Office of Management and Budget guidelines for U.S surveys (1978) specify an effort that would yield a 75% response rate. Other instances specify the original call and up to four recalls.

In a plan suggested by Kish and Hess (1959) sample points which yielded not at homes on former surveys become replacements for nonresponse contacts in a current survey. It is particularly suited to survey organizations that use similar sampling procedures. If replacement contacts receive the same number of minimum calls as the new survey, k -calls, the effect resembles a $2k$ -call procedure. The effect of non-response will not be completely eliminated since the "hardest-to-get" nonresponses may differ from the responses. The non responses from any current survey must also be similar to the nonresponses from the earlier survey from which the replacements are taken. Thus the time between the surveys must be brief enough to be reasonably sure that respondents have not moved or changed their attitudes and/or characteristics. Revisiting former refusals may also be a delicate task for interviewers to administer, though recent evidence collected in Lievesley (1986) suggests that a lot of refusals are situational - the result of unfortunate time of call rather than a perpetual sense of non-compliance. Lievesley suggests that around 30-40% of refusers can be persuaded to participate following a re-issue.

In Britain SCPR policy is for interviewers to make unlimited attempts so long as a recall can be placed conveniently on a workload route.

The necessity for repeated call backs to ensure the successful implementation of rigorous sample design may well be seen as a deterrent for probability sampling. If bias is to be avoided further calls must be made or a suitable method of adjustment developed. Table 9.10 (from Kish, 1965) shows it is exceptional to find more than 33 per cent of a sample at home on the first call; more recent data from Lievesley (1986) suggests this is still about the norm for first call productivity. (see tables 9.11 through 9.13).

Of course, as Frankel and Dukta (1983) point out if there is no relationship between the probability of obtaining a response (as measured by the difficulty of obtaining the response, like the number of calls made) then, there is no necessity to achieve a high response rate!

TABLE 9.10: Characteristics on several calls of a sample randomly selected adults from Dwellings in the Detroit Metropolitan Area

Demographic Characteristic	Number of cases	Percentage Distribution by Number of Calls					Mean No. of Calls
		One	Two	Three	Four or more	Total	
All Respondents	2213	32	30	14	24	100	2.6
Employment Status							
Employed	1357	21	31	18	30	100	3.0
Not employed	952	48	28	9	15	100	2.1
Age							
21-29 years	508	34	28	14	24	100	2.5
30-39 years	634	28	33	14	25	100	2.7
40-49 years	455	29	34	15	22	100	2.6
50-64 years	502	32	25	16	27	100	2.8
65 years or older	212	47	25	11	17	100	2.2
Marital Status							
Never married	186	28	21	14	37	100	3.3
Divorced, separated	105	26	24	16	34	100	3.2
Widowed	163	45	23	11	21	100	2.3
Married, no children present	864	31	26	16	27	100	2.8
Married, children present	990	33	36	13	18	100	2.3
Sex							
Male	1088	25	27	20	28	100	2.8
Female	1225	38	30	12	20	100	2.4
Relationship of Head of Household							
Head	1176	26	29	17	28	100	2.9
Wife	910	40	31	11	18	100	2.3
Other relative	189	33	31	14	22	100	2.6

Responses were 87 percent; refusals 8, not-at-home, and others 1 percent. Three Detroit Area Studies of 1957, 1958, and 1959 were combined by Sharp and Feldt [1959].

TABLE 9.11: Distribution of interviews by number of calls needed to achieve them

Result of Call	Percentage Distribution by Number of Calls						Mean No. of Calls
	One	Two	Three	Four	Five or more	Total Sample	
Completed interview	28	36	28	31	25	87	2.6
Refusal or "too busy"	20	16	21	15	18	8	4.6
Respondent not home	20	14	15	16	13	3	6.4
No one home	30	32	34	37	43	1	7.0
Respondent ill, senile	1	1	1	1	1	1	3.2
Language problem	1	1	1	*	*	*	3.7
Total	100	100	100	100	100	100	2.9
Number of cases	2646	1888	1164	767	1278	2651	7743

* Less than 0.5 percent.

Taken from Lievesley, 1986

TABLE 9.12: Distribution of interviews by the number of calls needed to achieve them

call no.	interviews %
1	30
2	32
3	19
4	10
5	5
6	3
7 or more	1
<hr/> 100 (based on 8029 interviews)	

(Taken from Lievesley (1986))

TABLE 9.13 Main categories of outcome by the number of calls made by the interviewer

call no..	interview %	recall arranged %	refusal %	non-contact %	established ineligible %	base %
1	19	23	5	36	2	100 (12549)
2	30	18	9	31	1	100 (8572)
3	30	16	9	32	1	100 (4955)
4	28	13	10	33	-	100 (2805)
5	26	13	10	39	1	100 (1508)
6	29	9	13	38	-	100 (738)
7 or more	30	12	13	47	-	100 (403)

(Taken from Lievesley (1986))

Kish's table (9.10) is useful in that it casts light on the possibility that the percentage distribution of certain respondent characteristics may remain fairly stable across calls whilst others will not. Again useful summary work from SCPR illustrates this phenomena and routine presentation of such data may eventually permit more informed decisions about call back strategies.

TABLE 9.14: Profile of sample for 1985 SCPR social attitudes survey by the number of calls made

	After 1 call	After 2 calls	After 3 calls	After 4 calls	After 5 calls	Attend of fieldwork
Owner Occupiers	52	58	60	61	62	62
Widowed	15	11	10	9	9	9
Church of England	38	38	36	37	37	36
Has no Qualifications	57	49	47	46	46	45
Conservative	26	30	30	30	30	31
Labour	41	37	37	37	37	36
Male	34	39	43	44	45	46
Looking after Home	30	27	23	21	20	19
Never had a job	14	10	8	8	7	7
Social Class 1 or 2	16	19	21	21	22	22
All respondents						

(Taken from Lievesley (1986))

According to Kish (1965) a plan or strategy for call backs should be included in computing costs, sample design and field procedure. Schedules should routinely carry accounts of timing and outcome of each call. Workloads should be ascertained not only on first call but also for call backs. Indeed the number of call backs need not be the same over the entire sample e.g with estimates of non-response by primary sampling units/areas different numbers of calls can be fixed in advance for different areas.

9.4: Data processing techniques for lessening the effect of nonresponse: an operational review of some approaches

9.4.1: Adjustments for bias without repeated call backs: the indirect approach

(a) Politz Simmons

This ingenious methodology was first suggested by Hartley (1946) and subsequently operationalized by Politz and Simmons (1949, 1950) and Simmons (1954) rests on the principal assumption that each individual in the population has a probability of responding if selected under the prevailing survey conditions. It is also implicit in a response:nonresponse model proposed by Platek et al (1977). The Politz plan includes questions to ask of each person found at home on the first call, and who does not refuse, whether s/he was at home last night at the same time, the night before last, etc., to cover the 5 nights preceding the interview, 6 nights in all. Each response is given a weight w_i , the reciprocal of the number of nights at home over the period of successive nights. Hence applying the Politz plan will produce the random variable.

$$x(P) = \frac{\sum w_i R_i x_i}{\sum w_i R_i} \quad (9.2)$$

where Σ denotes the sum over the six Politz classes, wherein, R_i and x_i denote the number of responses and their mean value in the Politz class. $W_i = 6 / (1+j)$, where j is the number of nights at home during the preceding five nights. W_i , R_i , and x_i are all random variables.

Comparisons on simulated populations by Cochran (1953), Deming (1953) and Durbin (1954) show this method to its best advantage, in relation to call backs, when biases from early calls are substantial and the sample is large. Obviously the method has the advantage of saving time but any errors in the values of i will be a considerable disadvantage. Simmons (1954) suggests that the method could be used in conjunction with repeated call-backs or as Deming (1953), in a survey interview plan where only temporary refusals require recalling.

(b) **Bartholomew**

The method proposed here is relatively straightforward: it involves eliminating all bias which distorts a survey estimate after the first call. Bartholomew (1961) advances both empirical and theoretical justifications for believing most, if not all bias arises at this stage. He suggests it is possible to obtain estimates which are practically unbiased after only two calls, the estimate for any survey characteristic being,

$$\bar{y} = (n_1/n) \bar{y}_1 + (1 - n_1/n) \bar{y}_2 \quad (9.3)$$

where n represents the number of calls on eligible contacts at the first attempt and n_1 the number of successful first calls, \bar{y}_1 and \bar{y}_2 represent first and second call means respectively. The method is useful if one has grounds for believing that the mean of the second calls is close to the non-respondent mean. Essentially the underlying assumption is that the successful second calls are a random sample of all respondents found to be not at home at the first call. The grounds for believing such an assumption depend on the interviewer being fully informed of the times each respondent was not at home so that s/he would be able to plan a route so that each member of the sample has approximately the same chance of being contacted. In practice the argument depends on interviewers obtaining information about the habits of potential respondents before making a second call. Neighbours or other members of the household may be useful providers of such information. Table 9.15 below provides evidence for the appropriateness of these assumptions.

In order to test the assumption that the second call will be a random sample of all respondents not found at the first call, we need to know what result would have been obtained if recalling had continued. For this reason, the sex of the respondent was used as the attribute under investigation. This was determined from first names on the Electoral Register for four random samples of electors. In each case the percentage of men in the first call is less than 100P by more than twice its standard error. Secondly, there was close agreement between the observed and expected percentages on the second call. (only for Ward A was the difference more than one standard error).

Bartholomew also presents information taken from Durbin and Stuart (1954), for appointment versus non-appointment calls at the second attempt. Table 9.16 suggests that the sex bias, present at the first call, was eliminated for second call appointments, whereas in the case of non-appointments the difference is significant at the 10% level.

TABLE 9.15: Comparison of percentages of men obtained at the first and second calls with those expected on the hypothesis of random sampling (Local Government Survey)

Ward	A	B	C	D
% of Men in First Call Sample	42.3 (52) 4.53	32.3 (62) 4.61	31.1 (61) 4.79	35.4 (48) 5.75
% of Men in Whole Sample (100P)	52.2 (90)	48.5(130)	49.6(137)	5.3 (130)
% of Men in Second Call Sample	59.3 (27) 4.98	61.4 (44) 4.35	66.7 (51) 3.87	63.8 (47) 4.65
% of Men left after First call (100P _r)	65.7 (38)	63.2 (68)	64.5 (76)	62.2 (82)

Note: standard errors for percentages are entered under percentages in rows one and three above.

TABLE 9.16: Percentage of successes at the second call in an experimental survey: total frequencies are given in parentheses (Durbin and Stuart's data)

Appointment Made		No Appointment	
Men	Women	Men	Women
72.5	69.0	34.6	44.3
(258)	(200)	(188)	(212)

(Taken from Bartholomew (1961))

Clearly the method will only be applicable if information can be obtained about non-contacts at the first call or if appointments are made. It would be unsuitable for mail questionnaires.

These procedures depend heavily on diligent interviewers carrying out their task in a careful manner. All three are perhaps most useful when conducted by fairly large scale survey organizations.

9.4.2: Post hoc imputation procedures which take no account of call back policy: passive approach

Any adjustment for incomplete data at the processing stage can be considered as a form of substitution. Non-response is only one (typically major) component of incompleteness, it can refer to total (questionnaire, unit) nonresponse or only to some questions on a schedule (partial or item nonresponse). Thus any review of imputation procedures will reflect methodologies that have been developed for varying aspects of incompleteness and types of nonresponse. Adjustment procedures may well be questionable as methods can have a substantial effect on the values and any resultant biases of survey estimates. So care is needed in appraising the impact of any procedure. The author is aware that there is an extensive literature on incomplete data, see Rubin (1987) whose text together with readings from the National Academy of Sciences Volumes on Incomplete Data (Madow et al. 1983) could adequately serve as a basis for a graduate course on imputation techniques. The review does not attempt to be exhaustive, but to put the analysis of incomplete data in the context of nonresponse in sample surveys.

Platek and Gray (1979) present four methods of imputation (zero substitution, weighting, duplication and the use of historical data), develop formulae for bias and variance with an application for a two stage sample design. Consider their general weighting formula provided overleaf:

General Estimation Formula for Estimates by Imputation Procedures

The estimation formula at the balancing area level b may be written in general by:

$$X_b = \sum_{i=1}^{n_b} (\delta_i w_i x_i + \delta'_i w'_i x'_i) \Pi_i^{-1} \quad (9.4)$$

where n_b = sample size in balancing area b

δ_i = event of responding or not responding pertaining to unit i
= 1 or 0 respectively

δ'_i = $(1 - \delta_i) \delta''_i$ = event of non-responding and availability or non-availability of historical records = 1 or 0, given that $\delta_i = 0$

δ''_i = event of availability or non-availability of historical records for imputation purposes

w_i = weight applied to responding units to enlarge the deficient sample resulting from non-response and/or lack of historical records

w'_i = weight applied to non-responding units with historical data available to enlarge the deficient sample resulting from non-response and lack of historical records

$x_i = X_i + R\epsilon_i$, when unit i responds; i.e., when $\delta_i = 1$

$x'_i = X_i + N\epsilon'_i$, when historical records are available and are substituted when unit i does not respond, $\delta_i = 1$

Π_i = probability of selection of unit i

Balancing areas are clusters of individuals with similar characteristics to the missing individual. i.e one is essentially replacing missing units by values that one would hope are close to the true value on the basis of partial knowledge of the nonrespondent. For total nonresponse of a unit the only knowledge typically available will be geographical location. Either a unit is allocated at random from a similar area as small as possible or the average characteristic of units in the balancing area are conferred on the missing unit. In practice there is a difficult balance to strike between setting up areas that are too large or too small, where they may be insufficient units to choose from (or high localized non-response). Clearly there will be an element of subjective choice in the selection of areas.

Where there is partial information available for a respondent post-strata can be defined for imputation purposes. These post-strata are referred to as weighting classes. Again the decision as to the best size post-strata is often a compromise of what is methodologically desirable and practically feasible.

In any balancing or weighting area there will be:

n_b sampled units

$n_b - m_b$ missing units, i.e. m_b respond, where m_b' may contain historical or source data.

Any deficiency in the sample caused by non-response is adjusted or enlarged by inflating the inverse selection probability (the "weight") by the inverse of the response rate. The inflation factor being

$(m_b + m_b')/ n_b$ in any balancing area b (9.5)

Not surprisingly this procedure is called the "weighting" method of imputation. Alternatives follow; one could do nothing of the sort, hence the "zero" substitution method, one could duplicate records of responders at random or use some form of historical substitution followed by reweighting. Platek and Gray present expressions for bias (see table 2 text) for survey estimates. Bias estimates are obtained over three stages:

- (i) over all possible responses
- (ii) over all possible subsamples of respondents, and
- (iii) over all possible samples

For a given set of response probabilities the bias of greatest magnitude occurs in the zero substitution method, otherwise the only other generalizable observation by the authors is that one would expect lower imputation bias when historical data are substituted as opposed to application of the weighting method alone (the duplication method has the same bias as the weighting method).

The duplication method described above is a special case of the application of the hot-deck procedure for imputing missing items as surveyed by Chapman (1976). Hot-deck procedures simply utilize information from current responders rather than cold-deck procedures which use information from past surveys. Pritzker, Ogus and Hansen (1965) applied the method in the 1960 U.S Census by substituting for a non-responding household the questionnaire responses of the previously listed responding household. In fact the idea was first discussed by Hansen, Huwitz and Madow (1953) who showed that the maximum increase in variance due to reweighting would be about 12%. Other extensions of the hot-deck procedure include the use of multiple regression or the AID program to define weighting classes (clearly designed with partial non-response in mind). For further illustrations of application the reader is referred to Chapman (1976) and Cox (1980).

Rubin (1979) presents a multiple imputation procedure based on Bayesian inference; he argues that in practice at least two imputations should be made under different imputation models thus allowing the assessment of sensitivity of inferences to different imputation models. For an interesting illustration refer to Shapiro and Schevren (1979).

9.4.3: Adjustments which take account of call-back policy: the direct/analytical approach

Deming (1953) was responsible for quantifying the nonresponse problem in a way to permit careful examination of the consequences of different call-back strategies. In his formulation we do not have two distinct strata of respondents and nonrespondents but the probability of response and non-response from each of several classes, each possessing a mean and variance. The classes range from an impregnable "hard-core" of no possible response up to a class of complete response. Deming identified six classes labelled 0,1,2,4,6,8 to denote the number of interviews completed on average out of 8 attempts. "0" denotes the hard-core or permanent refusal; "1 through 6" denote the temporary refusal, in the sense that an interview might be obtained with a different time of call or selection or interviewer. "8" represents the "soft shell" of respondents who are home 8 times out of 8 and always furnish a complete response. To define the likely bias arising from a particular attempt or call-back policy Deming uses the concept of a "patient mean", to imply the value attainable after patient recalling (ad infinitum) on people in classes "1 through 8".

The patient mean is a^* is thus:

$$a^* = \frac{\sum_{i=1}^8 p_i a_i}{\sum_{i=1}^8 p_i} \quad p_i = \text{proportion in class } i, a_i \text{ mean for class} \quad (9.6)$$

Summing across all classes including "0" is permissible as they yield no information, but clearly the patient mean conditions on those ever likely to respond and in some sense represents a truncation of the survey population. The expression for the resultant bias is:

$$B(I) = E(I) - a^* \text{, where } I \text{ refers to attempt} \quad (9.7)$$

As class 8 is wiped out in the first attempt the bias of non-response arises from the absence of classes "1 to 6"; successive calls are assumed to dig deeper in lower classes diminishing the relative proportions that remain in the upper classes thus pushing the accumulated result nearer to the patient mean. Deming assumes

each attempt picks up a random sample of non-respondents in each class. This is equivalent to Bartholomew's assumption only if interviewers are able to obtain information or cooperation from all classes. Deming also shows that the model produces a similar bias term to the Politz plan under $k=2$ recalls and modest algebraic simplifications.

Deming's results suggest that without recalls it is hazardous to put confidence in the result based on the initial calls and with an estimate formed by summing the initial call and the first two recalls effect together about a 50% reduction in the nonresponse bias. Even increasing recalls to six enhances the amount of information obtained per dollar. Analysing call-back means therefore provides interesting insights into the effects of nonresponse. Deming's model provides an important and flexible framework to assess call-back strategies and a basis for developing probabilistic models. It anchors the remainder of the review.

An early example of using call-backs or reminder mailings as indicators of the amount of effort required to obtain a response was carried out by Hochstim (1967) in the context of a mail survey, the procedure is known as the "**successive stages model**". This model assumes a relationship between the difficulty of obtaining a response, judged by the number of mailings and/or call-backs or rediallings to obtain a response and a given variable. Linear extrapolation beyond the number of attempts employed is possible, using least squares estimation. Locker, Wiggins et al (1981) applied the method to a mail screening survey to estimate the prevalence of disability. The method is considered justifiable where cumulative responses by attempt show a linear trend (Hochstim and Athanasopoulos, (1970). More recently McGowan (1986) has demonstrated a procedure of fitting S-shaped curves to mail response data. Chapman (1976) discusses the idea of using call-backs to identify weighting classes in the general weighting method of imputation discussed in the previous section. He questions the validity of the assumption that the survey characteristics of the nonresponders will be more alike the late cooperators than those for all responders. An intensive follow up of CPS nonrespondents by Waksberg and Pearl (1965) indicated little support for the assumption. Chapman also investigated data collected in the Health and Nutrition Examination survey (1974)

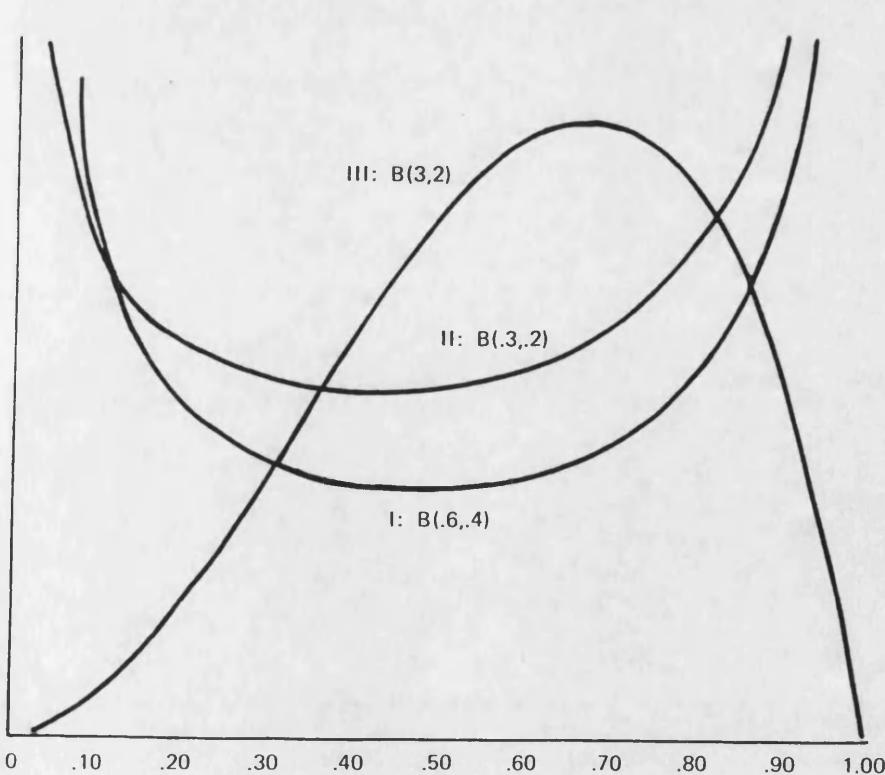
and found many different patterns for various survey items; a general trend was not apparent. His conclusions remain somewhat pessimistic in that even where a trend is apparent he questions the existence of an appropriate method of extrapolation.

Despite such caution various authors have persisted with the construction of models based on the assumption that success in obtaining an interview is a probability process based on repeated calls. Frankel and Dutka (1983) extend Deming's original formulation of establishing a finite number of response probability classes to a model which postulates the existence of a latent continuous-probability function, $f(p)$. This density function of response probabilities is in the form of a beta distribution as

$$f(p) = p^{u-1} (1-p)^{v-1} \quad u, v > 0 \quad 0 < p < 1 \quad (9.8)$$

For specific survey conditions, e.g. fieldwork agency used, different values of the parameters u and v exist. Three such functions all having the same average response probability are shown in the figure below:

Figure 9.1 Some latent response functions



The population being sampled can be represented by the area, A, of the density function so that if n calls are made to reach this population, the proportion reached can be expressed as

$$\frac{A_n = \int_0^1 f(p) (1 - (1-p)^n) dp}{A} = \frac{B(u, v) - B(u, v+n)}{B(u, v)} \quad (9.9)$$

Table 9.17, from Frankel and Dukta shows the percentage of the sample reached as the number of calls increases from one to ten. For example, an 80% coverage would be obtained by the fourth call for distribution I, the ninth for distribution II and the second for distribution III. The differences for each of these distributions is summarized in table 9.18. Under infinite recalling the overall probability of being interviewed is the same, ie. .60. However, given different latent response probabilities for respondents one can demonstrate the operation of repeated calling upon a sample. For the same coverage, say 80%, the average response probabilities achieved for distributions I-III are .71., .74 and .64 respectively. Frankel and Dukta show even with an 80% coverage achieved after four calls, those individuals with less than 30% probability of responding are highly under represented. The important point arises as to what effect these differences in underlying response-probability distributions will have on some variable being estimated. If there happens to be a correlation between the value of the variable and the probability of a response the effect could be considerable.

To assess these implications two separate linear functions were selected and constructed, both being related to the probability of response and both having average values that were the same for the three distributions considered. The results demonstrated that it is not only necessary to take into account the percentage of the designated sample reached, the nature of the relationship of the variable being measured, and the response rate, but also the shape of the latent response function.

The authors also go to consider the case when there is concern about the costs of making call backs. In these situations, how much extra effort should be expended to reach the difficult cases? For the models considered, distribution II appeared hopeless.

When such a scenario emerges they indicate that it might be possible to redefine the target population (e.g market research practices to eliminate persons living in poverty areas of cities would be, regarded as 'low probability of response at high cost')! In modelling terms this would mean truncating the distribution of response probabilities. Frankel and Dukta claim that their modelling procedures have provided guidance in the design of new studies and refinement of existing continuous surveys. Although, exact values of the parameters for the beta functions are unknown they have a general idea of the response curves for the various population groups under study. The practice has helped determine call back strategies, when to resort to the use of incentives, when to shift to different modes of interviewing, and when to shift to exotic techniques should repeated calls fail!

TABLE 9.17: Proportion of sample reached by repeated calls (R)

Number of calls	DISTRIBUTION		
	I (.6,.4)	II (.3,.2)	III (3,2)
Infinite	1.00	1.00	1.00
1	.60	.60	.60
2	.72	.68	.80
3	.78	.72	.89
4	.81	.74	.93
5	.83	.76	.95
6	.85	.77	.97
7	.86	.78	.98
8	.87	.79	.98
9	.88	.80	.99
10	.89	.81	.99

(Taken from Frankel and Dukta)

TABLE 9.18: Average response probabilities of those reached (P_r)

Number of calls	DISTRIBUTION		
	I (.6,.4)	II (.3,.2)	III (3,2)
Infinite	.60	.60	.60
1	.80	.87	.67
2	.76	.83	.64
3	.73	.80	.63
4	.71	.79	.62
5	.70	.77	.62
6	.69	.76	.61
7	.68	.76	.61
8	.68	.75	.61
9	.67	.74	.61
10	.67	.74	.60

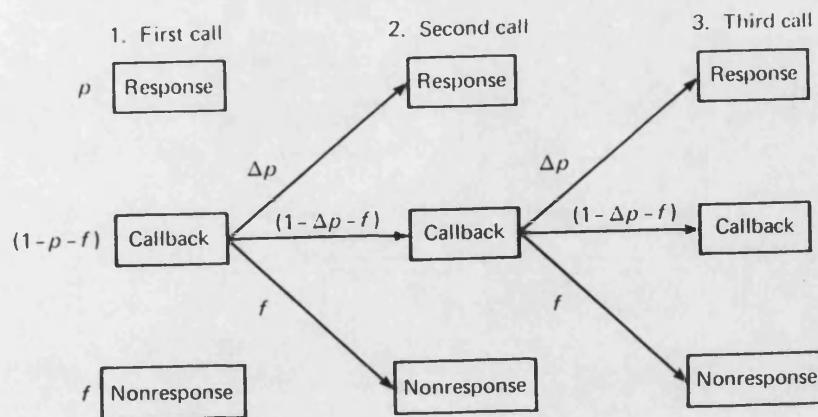
(taken from Frankel and Dukta)

Thomsen and Siring (1983) propose a probabilistic model that also considers availability to be important when correcting for nonresponse bias. They illustrate two methods of estimating parameters for situations where there are two poststrata (here age groups); one where the number of persons in each poststrata is known and the other where they are not known. In the first instance maximum likelihood estimates can be calculated and in the latter estimates are calculated using least squares. For computational convenience the methods are applied to the same survey. Obviously, in practice only one method would be used. Models are fitted to data from the Norwegian Fertility survey where the objective considered is the estimation of the mean number of live births to women in the population. The model classifies attempts to obtain an interview in three ways:

- (i) successful response
- (ii) no response interviewer decides to call back
- (iii) ditto (ii) but categorize as refusal

The authors suggest that in practice it is difficult to distinguish between temporary and permanent refusals. If p denotes the probability of outcome (i) and f the probability of outcome (iii), both at the first call, then $(1-p-f)$ must denote the probability of outcome (ii). "f" is assumed to be constant for all calls, outcome (i) is designated probability Δp for subsequent calls, where Δ is typically expected to be >1 to indicate interviewer ingenuity regards respondent availability for recall (planned timing and/or appointments). Figure 9.2 below summarizes the framework of the model:

Figure 9.2: A probabilistic model for nonresponse



Assuming that the parameters p, f and Δ are constant within the two poststrata maximum likelihood estimates are developed. The method enables conditional probabilities of women being categorized as nonrespondents at each attempt to be estimated. In their application for the younger strata (18-29 years) the observed conditional probabilities show a clear tendency to decrease. The authors suggest this may be due to the possibility that later calls are made on people are difficult to find or the tendency of interviews to categorize a respondent as a refusal instead of deciding to call back. This observation could be accommodated by allowing for a shift in the value of f and/or p within poststrata (for example assuming p is generated by a beta distribution). However the authors retain the simplicity of the original specification on the basis of a satisfactory goodness of fit.

In the instance where the numbers in poststrata are not known, the method of least squares is applied to obtain estimates. Essentially p and Δ are again assumed to be constant within each strata but vary between them; in addition, f was assumed constant for the whole sample. Assuming N women have been selected in the sample and that there are N_i in poststratum i , then the expected number of responses in the j th call in poststratum i is $N_i P(C_i=j)$, where

$$P\{C_i = c_i\} = \begin{cases} p_i & \text{if } c_i=1 \\ (1-p_i-f) (1-\Delta_i p_i - f)^{c_i-2} \Delta_i p_i & \text{if } c_i \geq 2 \end{cases} \quad (9.10)$$

If X_{ij} denotes the observed number of responses in poststratum i after the j th call then least squares estimates of Δ_i , p_i and f can be found by minimizing

$$\sum_{i=0}^6 \sum_{j=1}^8 (N_i P(C_i = j) - X_{ij})^2 \quad (9.11)$$

under the condition that $\sum_{i=0}^6 N_i = N$. Thomsen and Siring reduce the number of parameters by assuming $p_i = i\beta + \alpha + \text{residual}$. Using data from the first three calls for seven poststrata demoting the number of live births to women, ranging from 0 to 6, they show a 40% reduction in the bias of the average number of live births based on an adjusted average using estimated values of N_i .

Proctor (1977) simplifies Deming's model to develop a maximum likelihood estimation procedure to estimate a proportion in a survey with call-backs; although he suggests his results may be more appropriate to telephone interviewing. Responses are considered as zero/ones respectively with " α " and " β " chances of not responding. Short of infinite recalling ($r = \infty$) Proctor shows there will be a bias in the survey estimate of the population proportion P (where $Q = 1 - P$) equal to

$$PQ(\alpha^r - \beta^r)(1 - Q\alpha^r - P\beta^r)^{-1} \quad (9.12)$$

Following Deming's supposition that after r calls the observed frequencies that answer zero or one follow a multinomial distribution it is possible to obtain by Rao's method of scoring (1952) maximum likelihood estimates of α , β and P . Clearly, after r calls there will be a residual frequency (n_{r+}) of nonresponders. In applying the methodology this value is reset so as to obtain coincidence of observed and fitted values of zeroes and ones (refer to tables 1 and 2 in the original text). Any difference between the fitted and original value of n_{r+} is taken as an estimate of the

combined numbers of hardcore nonresponders and non-working telephones. Essentially the adjusted estimate of P is conditional, as it represents the proportion of ones after excluding the fitted value of n_{r+} .

Finally, within the realm of general probabilistic models considered so far is an interesting application due to Drew and Fuller (1980). For a simple random sample wr of n individuals from a population of N , it is supposed that the population is partitioned into k categories or strata corresponding to values of a discrete random variable. Associated with each individual of the k th category is a response probability (q_k) that the individual produces a complete response when sampled. For some $q_k = 1$, then $n_1 < n$ responses are obtained on the first call. The $n-n_1$ nonrespondents at the first call then represent the frame for the second call, where n_2 individuals are assumed to respond. Calls continue in this way until, after r calls n_{r+} individuals have not responded (equal to the residual frequency denoted by Proctor).

It is also assumed that a proportion of the population is composed of hardcore nonresponders who will never answer the survey ($=1-v$). In their initial treatment v is assumed to be constant in each category. n_{rk} the number of individuals responding from the k th category responding to the r th call and n_{r+} are observed from the survey. If the population proportions in the k categories are denoted by a vector f_k where $\sum f_k = 1$, then the data satisfies a multinomial model with response probabilities.

$$\begin{aligned}\Pi_{rk} &= v (1-q_{rk})^{r-1} q_k f_k & (9.13) \\ \Pi_o &= (1-v) + v \sum_k (1-q_k)^r f_k\end{aligned}$$

where Π_{rk} is the probability that an individual in category k responds on call r , and Π_o is the probability that an individual will not have responded after r calls. Vectors f_k and q_k and scalar v are unknowns and can be estimated by maximising the log likelihood, $\log (n:f,q,v)$ which is proportional to

$$\sum^r \sum^k n_{rk} \log \Pi_{rk} + n_{r+} \log \Pi_o \quad (9.14)$$

The likelihood maximization is achieved by the method of scoring (see Rao, 1952 and 1973). The model is based on the assumption that the response probabilities depend on the number of calls, k . As many survey variables are continuous Drew and Fuller also consider grouping data on the basis of a continuous variable.

In the discretized case they show how estimates of f_k can be used to construct estimates of means of other variables in a survey by applying (9.14) to a postal survey of households where three mailings were made to an initial sample of 1023. The sample respondents were divided into seven age categories. Further inspection of the estimates obtained for q_k led to a reparameterization that specified a quadratic relationship between q_k and the median age in each category. Alternative assumptions relating to the allocation of hardcore nonrespondents across categories and the degree of intensity of call backs were also introduced (relevant for this application because of differences in style and content of reminders). All formulations provided acceptable goodness of fit.

Testing different imputations involving varying assumptions is clearly valuable. If the assumptions reflect a realistic assessment of the context of data production then they will help analysts understand the process of nonresponse. In the illustration that follows good information on call-backs and appointment procedures together with appropriate software (Maximum Likelihood Program, Ross 1983) made Drew and Fuller's approach realisable.

Not all methodologies are software dependant. For instance Bartholomew's procedure, can be adapted to refine a survey estimate by assuming that any outstanding potential interviews remaining after k calls have a mean equal to that generated by respondents at the k th stage.

These two procedures will be demonstrated in an evaluation of a call back strategy where the cost of obtaining information is considered important. First, it is important to establish an evaluation framework that makes a sensible connection between costs and potential bias or distortion in the survey estimates. This follows in the next chapter.

Chapter 10:

The interplay of non response bias, costs and other sources of error

Contents

- 10.1 Survey costs and mean square error-an overview**
- 10.2 The development of cost functions**
 - 10.2.1 Simple cost functions**
 - 10.2.2 More general cost functions; particular reference to those that take account of repeated attempts to obtain information**
- 10.3 Two illustrative examples of the interplay of costs and design**
 - 10.3.1 An introduction to the illustrations**
 - 10.3.2 Sudman's work on costs of probability samples with call-backs and quotas**
 - 10.3.3 Durbin and Stuart's experimental study on call-backs and clustering in sample surveys**
- 10.4 The search for optimal design**
 - 10.4.1 Efficient design criteria**
 - 10.4.2 Multipurpose allocation**

10.1:

Survey costs and mean square error: an overview

The two primary considerations in survey design are costs and precision. Ideally, researchers would like to minimise cost and maximise precision. Typically, survey budgets are 'set' or 'fixed' so we attempt to maximise precision for a given cost.

Precision is a term usually reserved to the size of deviations from the mean m obtained by repeated application of the sampling procedure (Cochran, 1963 p.16). It is preferred to the use of the term accuracy whenever we lack confidence as to the presence of unsuspected bias in our estimates. Accuracy refers to the size of deviations from the true mean. Inclusion of a bias factor in the initial statement simply implies substituting the term accuracy for the term precision. As introduced in Chapter 1 the criterion "mean square error" will be used to convey the notion of accuracy of a survey estimate. Methodologies reviewed in chapter 9 enabled at least one source of bias, namely nonresponse bias to be estimated, so as to approximate mean square error.

The survey designer is often without cost information. The survey used in the illustration that follows is no exception to this norm. Decisions must be based on estimates or guesses. To help make good guesses or estimates practitioners have to rely on good cost models. A number of models are reviewed in the following section.

Two illustrative examples from the survey literature are presented to provide the reader with a pragmatic understanding of the need to assess the interplay of costs and design. In the final section a global criterion for survey design evaluation is proposed which reflects earlier work by Deming (1953), Durbin (1954), Durbin and Stuart (1954) and Kish (1965) in the context of Kish's more recent work on optimal/proximal solutions for multipurpose surveys.

10.2: The development of cost functions

10.2.1: Simple cost functions

For illustrative purposes it is useful to consider how costs enter a survey involving the use of cluster sampling. Ordinarily cluster sampling results in less precision than a sample of individuals selected srs wr. On the other hand cost per sampled individual is lower. In theory cluster sampling should be used whenever its effect on costs is greater than any decrease in precision. Hansen, Hurwitz and Madow (1953, Vol I, p. 270) identify three main cost components of the various phases of a survey:

- (i) fixed costs or overheads; costs of central administration, technical support etc., which may be assumed to be the same even for marked variations in size and design. In this sense the total cost considered refers to "total variable costs", made up of (ii) and (iii) below.
- (ii) costs that vary in proportion to the number of primary units (clusters) in the survey. These costs include costs of selection, travel to and location of psu's, and any necessary subsampling (and/or frame preparation) within psu's. Cost per unit = C_1 .
- (iii) costs that vary in proportion to the number of listing units (individual units sampled) in the sample; included here will be the direct costs of interviewing and analysis share. Cost per individual unit = C_2 .

The total survey cost can be expressed as:

$$C = C_1 m + C_2 m \bar{n} \quad (10.1)$$

where C is exclusive of fixed overhead, where m is the number of primary units in the sample and \bar{n} the expected subsample size per primary unit in the sample.

This model is essentially the same as the functions suggested by Cochran in the exploration of how cost depends on the number of strata used in stratified random sampling (1953, p. 134) where m would represent the number of strata and mn the number of individuals sampled. Sukhatme and Sukhatme (1970) also construct a parallel model in their initial examination of survey costs.

Adopting this style of cost function Hansen, Hurwitz and Madow proceed with an illustration of a method for determining the optimum number of interviewers in a survey (under fixed essential conditions). Here (Vol II, p294) C is again exclusive of fixed overheads and

$$C = C_1 n + C_2 m \quad (10.2)$$

for n sampled individuals and m interviewers. C_1 then is the cost per individual (presumed to be direct costs of interviewing and sampling) and C_2 the costs per interviewer (presumed to represent indirect costs of training, briefing, supervision and travel). The authors suggest that for a fixed total budget increasing the number of interviewers will increase costs and require a reduction of costs at some other point. e.g expenditure per interviewer or reduction in sample size. Equipped with this cost function and appropriate expression for the variance they demonstrate how optimum values for m and n can be determined for a joint solution of the cost and variance functions (p. 295, Vol II). The methodology is illustrated for two surveys conforming to design specifications in the Mahalanobis tradition. Without actually defining the term their conclusions indicate "proximality", a state of being satisfactorily within some mathematically desirable optimum, or best possible state e.g. "for m between 4 and 16 the variance of the sample mean will be within 13 per cent of the optimum".

Their work is also important because they raise the question that in choosing a single sample design the investigator may have a choice of alternative methods, ... "each with different essential conditions, response bias, and optimum values for m and n ". In particular where it is evident that a particular survey technique is subject to substantial response bias, alternative

techniques may be available to reduce the bias but may be relatively expensive. It is in this context that "double sampling" schemes are useful, first interview or collect information for a large sample by a cheaper less accurate method, followed by a reinterview or indepth technique for a subsample by the more expensive method. This procedure is aptly demonstrated in three illustrations (Vol II, p. 303/4).

These formulations demonstrate the need when planning surveys to consider a general expression to compare alternative designs. Kish (1965) provides such a vehicle. As all of the formulations considered so far the formulation may only serve as a rough approximation to complex reality but it helps to simplify an otherwise daunting process. Here total costs, both fixed and variable are subsumed under one heading, T., such that

$$T = K + C \quad (10.3)$$

where C could be considered as representing the earlier Hansen, Hurwitz and Madow formulation. However a subscript v is introduced to convey the notion that elements of fixed costs may vary with the type of design, for instance the cost difference in computing variances for cluster sampling versus srs wr. More obviously perhaps, variable costs are allowed to differ according to design differences, e.g. costs of locating fieldworkers, travel etc., Thus for n respondents,

$$T = K + K_v + nc + nc_v \quad (10.4)$$

Thus the total cost of a survey is expressed as the sum of four component classes each of which may contain several distinct factors. Following Kish (1965), K is the class of constant cost factors, which do not change either with the number of sample elements nor with the type of design used. K_v denotes the class of cost factors which vary with changes in design, but not the number of sample elements. nc denotes the total class of factors which are proportional to the number n of sample elements, which are not affected by changes in the sample design. nc_v denotes the total cost of factors that are proportional to the number of

sample elements, but also vary with changes in design. Kish notes that c_v and c need only be averages: since only a few aspects of the design can be specified, others are taken on the average. The question of exactly how these average costs are defined or vary within the entire sample will be given more attention in the following subsection. For now the formulations permit illustration of important aspects of evaluation, namely the definition of indices to reflect the two principal considerations of design, cost and precision (accuracy). They also determine a foundation for consideration of how to attain 'optimal' designs. Kish indicates that the relative advantages of designs can be expressed in terms of element (unit) cost and variance, if a bias factor is thought appropriate then read mean square error for variance. The product of these two components can be thought of as a criteria of evaluation of design. A design is preferable when it has a smaller cost per unit variance or a smaller variance per unit cost. Consider the ratio:

$$\begin{aligned}
 \frac{\text{Cost}_v \times \text{Var}_v}{\text{Cost}_{v'} \times \text{Var}_{v'}} &= \frac{(\text{element cost} \times \text{element variance})_v}{(\text{element cost} \times \text{element variance})_{v'}} \\
 &= \frac{(c + c_v) \times \text{Deff}_v}{(c + c_{v'}) \times \text{Deff}_{v'}} \quad (10.5)
 \end{aligned}$$

for designs v and v' . Design v' will be preferable when the ratio > 1 , v will be preferable when the ratio < 1 .

In a manner similar to Hansen, Hurwitz and Madow, Kish presents a formula to designate the most economical subsample size for cluster sampling which serves to illustrate a general rule about optima. Essentially where unit variance may be expressed as a linear function $V = (w + W \bar{n})$ increasing with n (a consequence of clustering effect), and unit cost as a linear function $C = (c + C_1 / \bar{n})$. Then the product $V^2 C$ denotes variance for unit cost and differentiating yields a solution for n that defines a minimum for $V^2 C$. The implication is minimum variance for unit cost or minimum cost for unit variance.

10.2.2:

More general cost functions: with particular reference to those that take account of repeated attempts to obtain information

All of the authors mentioned in the previous subsection recognise the need to develop cost functions beyond the straight forward generalizations presented so far. However, there are reservations; whilst the introduction of more complexity may attend more closely to reality the cost may be the subject of unreliable estimation. In the author's view one outstanding aspect of design implementation that does deserve more explicit attention is the inclusion of a specific component to reflect the amount of effort often required to obtain a survey response. In 'face to face' interviews this will imply consideration of call back norms and planning as well as the size of a geographical area to be covered.

First, consider the simple cost function considered in the previous subsection. Hansen, Hurwitz and Madow show how travel costs may not be properly reflected. Travel costs are considered in three parts: travel costs between psu's, travel costs within psu's and travel costs from home or office to and from the psu's. With small psu's the authors consider travel costs to be relatively small and suggest that they are simply reflected in a cost element for sampling within a psu, focusing their argument on the other two components. Excluding a component in this manner may not be always appropriate. Clearly how small does the area of a psu have to be for the assumption to be permissible? Under certain simplifying assumptions Slater and Wiggins (1978) suggest allocation schemes do not always lead to such an obvious conclusion. To illustrate how travel costs vary depending on the number of psu's in the sample Hansen, Hurwitz and Madow examine the distance involved by beginning at one corner and travelling by the shortest possible route from one point to another, where points are equally spaced intervals throughout an area such as that indicated by Figure 10.1 overleaf.

Figure 10.1:

Distances between points of different densities in a rectangular area of 960 square miles.

(Taken from Hansen, Hurwitz and Madow, vol 1)

$$\text{Area} = A = 40 (24) = 960 \text{ sq. mi.}$$

-----	0	0	0	0	0	0	0	0	0	0	0
$d_1 = 4 \text{ mi}$	0	0	0	0	0	0	0	0	0	0	0
-----	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

$$m_1 = 60 \text{ points}$$

$$d_1 = \sqrt{A/m_1} = 4 \text{ mi}$$

-----	0	0	0	0	0
$d_2 = 8 \text{ mi}$	0	0	0	0	0
-----	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

$$m_2 = 15 \text{ points}$$

$$d_2 = \sqrt{A/m_2} = 8 \text{ mi}$$

For m equally spaced points in the area, A it can be shown that the distance between any consecutive pair of points is:

$$d = \sqrt{\frac{A}{m}} \quad (10.6)$$

Thus the average distance between points is inversely related to the square root of the number of points to be visited. With m points to be visited the total distance between them by the shortest possible route will be $(m-1)d$ or approximately md , and since $d = \sqrt{A/m}$, approximately equal to \sqrt{mA} .

So far, it has only been assumed that one visit per psu is necessary. Assuming that all first calls were made before making any call backs, then the cost for making a second visit would be a function of the proportion of calls outstanding and the total distance to be travelled for all second calls. Proceeding in this way the additional cost factor that would need to be introduced into the original formulation is seen to be a term $Co\sqrt{m}$. Clearly assumptions regarding call back routes and productivity of calls need to be examined carefully. For the impact of travel from home, office to and from sample psu's Hansen, Hurwitz and Madow suggest adjusting the cost coefficients to take account of the estimated cost for beginning and ending trips. As a factor this increases the direct cost of time spent by interviewers in the field. This modification of approach is used to define a more complex cost function. Hansen, Hurwitz and Madow (1953, Vol II, p.173, 188 and 223) develop theorems and proofs using Lagrangian functions to find optimal design solutions for clustered sampling and two and three stage stratified designs.

Another approach to developing a more general cost function is to include call-backs or repeated attempts at securing information into the formulation. The clarification is due to Durbin (1954). He suggested that a common fallacy pervaded regarding the costs of call-backs. Investigators might be tempted by the following idea "each successful first call has cost an amount c . Each successful second call has cost $2c$, since two calls were made. Similarly, each successful third call has cost $3c$. Therefore, the longer

recalling is continued, the more expensive the interview becomes" ...Durbin goes on to point out that for successive calls 1..3 (say) costs of successful calls (interviews completed) will be $n_1c/s_1, n_2c/s_2, n_3c/s_3$ i.e they will depend on the relative success rates $s_1/n_1, s_2/n_2, s_3/n_3$ of achieving interviews.

Table 10.1 below taken from Kish shows the results for two cost models based on the explicit function $a_i s_i = c_i n_i + K s_i$ for the i th wave of calls providing differing values of a_i , the unit cost per call. Thus $\sum^k a_i s_i = C$ for k call-backs including the first call. a_i is equivalent to $K + c_i n_i/s_i$, where K represents the direct cost of interviewing. Summing across call-back waves implies $K \sum^k s_i$ is equivalent to the sum of all direct interviewing costs (analogous to ' $C_2 m\bar{n}$ ' in 10.1) plus a component which reflects the average variable costs of making calls at particular waves. Obviously the relatively more successful call-backs are the better (s_i will always be $\leq n_i$). Kish's formulation does not allow for travel costs between psu's.

TABLE 10.1:
**Two models of mean cumulated costs with
6 calls on not-at-homes**

			MODEL I OF COSTS			MODEL II OF COSTS		
Wave of Call	Responses	Cumulated Responses	Cost per Response	Cost for Wave	Mean cost of Cumul. Resp.	Cost per Resp.	Cost for Wave	Mean cost of Cumul. Resp.
(1)	(2)	(3) $\sum n_r$ $r=1$	(4)	(5)	(6)	(7)	(8)	(9)
1	42	42	1.0	42.0	1.00	1.0	42.0	1.00
2	35	77	0.9	31.5	0.95	1.1	38.5	1.04
3	14	91	1.1	15.4	0.98	1.4	29.6	1.21
4	4	95	1.4	5.6	1.00	2.0	8.0	1.24
5	2	97	2.0	4.0	1.02	3.0	6.0	1.28
6	1	98	2.4	2.4	1.03	4.0	4.0	1.31
TOTAL	98			100.9			128.1	

(taken from Kish, 1965)

In model I unit costs on the last three calls are higher because the cost per call and the calls per response are higher. However because their proportions are low they only have modest effects on the overall average variable cost. In model II unit costs for later calls are set to go up more sharply, as Kish argues this may represent a situation where travel costs rise sharply because the sample is spread, and the basic interview cost is relatively low.

In this sense overemphasis on travel costs "can mislead us to overestimate the cost increase for later calls", especially if the fixed cost per interview, K , is high relative to travel cost. He goes on to suggest that the first call cost should ... "include the entire cost of selecting sample cases, finding them, and identifying them.." a rather contrasting perspective to Hansen, Hurwitz and Madow who levy such costs on the sampling within a psu component. Accepting Kish's argument as a better reflection of the task facing fieldworkers would imply placing the cost of first calls at even greater disadvantage than is the case presented in the table above.

Using the model described by (Thomsen and Siring (1983)) in section 9.4.3 the authors study the relationship between mean square error and the number of call-backs. In particular, they examine the allocation of resources between the initial sample and the number of planned call-backs. They assume travel costs per visit are constant and unlike Kish suggest this is a departure from the common assumption made elsewhere that first call costs are less than the following calls, without substantiating their claim. Nevertheless their exploration is important for several reasons. Assuming the survey objective is to estimate the number of live births to women in the Norwegian Fertility survey for a fixed budget and at most eight calls, they demonstrate a declining mean square error with the number of call-backs (table 10.2 below), they implicitly define the first call logically as a "zero" call-back, denoting a commitment to at least one call on all potential respondents, and show under the various assumptions of their model it seems reasonable to select a relatively small sample and to use a large proportion of resources on call-backs. (refer to tables 10.3 to 10.5 below). This wisdom is echoed by Deming (1953, p.67, p69) on the futility of sheer size of sample to combat nonresponse.

TABLE 10.2:

**Estimated bias and mean square error by
number of calls**

Number of calls										
	1	2	3	4	5	6	7	8	9	10
Bias	.339	.207	.166	.144	.133	.127	.123	.122	.121	.120
Mean Square error	.1162	.0435	.0281	.0212	.0182	.0166	.0156	.0153	.0151	.0149
$\sqrt{\text{Mean Square error}}$.3408	.2085	.1676	.1457	.1347	.1288	.1248	.1238	.1228	.1219

TABLE 10.3:

Total costs (N.kr) by number of calls

Number of calls									
	1	2	3	4	5	6	7	8	9
TOTAL	151,410	243,120	285,510	305,460	314,970	319,560	312,870	322,980	323,520
COSTS (N.kr)									

TABLE 10.4:**Strategies that all cost N.kr. 322,980**

Strategy	Initial Sample	Number of Callbacks	Strategy	Initial Sample	Number of callbacks
1	10,766	0	7	5064	6
2	6,705	1	8	5047	7
3	5,709	2	9	5039	8
4	5,336	3	10	5034	9
5	5,175	4	11	5032	10
6	5,101	5			

TABLE 10.5:**Mean square error of the sample mean for the different strategies in table 10.4**

1	2	3	4	5	6	7	8	9	10	11
Mean Square error	.1155	.0433	.0280	.0212	.0181	.0166	.0156	.0153	.0151	.0149

(All tables taken from Thomsen and Siring, 1983).

10.3:

Two illustrative examples of the interplay of costs and design

10.3.1: An introduction to the illustrations

Careful examination of the ingredients of the cost models reviewed in section 10.2 necessarily involve close inspection of the aspects of conducting a survey and how they may affect costs. Such an exercise may provide both insight into the realism of the model and guidance as to how to identify sources of cost in a particular survey. Hansen, Hurwitz and Madow (1953) arrive at such a list by identifying the various tasks involved in implementing a survey. For the purposes of consolidation their listing is summarized below:

- (i) planning and broad direction of survey; typically office and field contributions to overheads.
- (ii) immediate field supervision of interviewers; may be considered as a proportionate increase on direct interviewing costs.
- (iii) selecting sample of psu's; part of the preparation of sampling frame could be seen as part of the overheads.
- (iv) direct payments to interviewers, made up as follows: training, travel between psu's and costs of listing/ and/or sampling within selected psu's. Most of these issues were considered in 10.2.
- (v) editing, coding and processing; could either be seen as part of overheads or as a proportionate increase on direct cost of interviewing.
- (vi) analysis and report writing; could be included in (v) of course.
- (vii) printing costs; typically a component of fixed overheads.

In carrying out such an exercise the investigator must be aware that the objectives are not simply bound with cost control but to realise optimal designs. Two illustrative examples follow (Sudman, 1961 and Durbin and Stuart, 1954). They are considered valuable because of the emphasis they give to costing in evaluation. Durbin and Stuart's application is particularly noteworthy as they provide a global evaluative framework for both design, in its broadest sense, and cost. Interestingly, both studies arose in the context of a practical debate between the relative merit and demerit of non-probability sampling (quotas) versus probability sampling. In order to achieve the full benefits of probability sampling (namely estimates of precision and, possibly, appreciation of the extent of unbiasedness) it is assumed that one is committed to "costly" call-back (repeated effort) procedures.

10.3.2:

Sudman's work on costs of probability samples with call-backs and with quotas

Sudman (1961) compares the costs of various National Opinion Research Center (NORC) probability call-back and quota samples, and indicates that a substantial portion of the cost differential them is not due to field activities but between other aspects of the studies unrelated to sampling. Six NORC probability call-back and four quota studies are compared. Table 10.6 taken from Sudman demonstrates for real cost information call-back costs per unit typically three times as high as the quota samples. More careful examination of the table reveals that a substantial part of this differential is due to differences in planning, processing and analysis. Almost always planning and analysis of call-back studies is costlier than quotas. In this manner it is not the sample design that determines the cost, but the cost that determines that design.

Table 10.6:
Field, and other costs for NORC surveys (Sudman, 1961)

COSTS	PROBABILITY WITH CALL-BACKS						PROBABILITY WITH QUOTAS			
	1	2	3	4	5	6	1	2	3	4
Direct Field Costs	\$31,800	\$21,000	\$19,500	\$5,000	\$22,000	\$16,000	\$8,900	\$9,900	\$8,500	\$9,000
Field Supervision	8,100	29,500	4,900	2,500	9,500	6,000	1,900	1,700	1,200	1,900
Other Survey Costs	173,100	106,200	93,400	31,400	38,500	26,500	16,000	18,000	14,100	14,800
Total Costs	213,000	156,700	117,800	38,900	70,000	49,400	26,800	30,200	23,200	25,700
Total Cases	2,380	2,810	2,200	760	2,500	1,500	1,200	1,500	1,300	1,500
Cost/Case	89.50	55.80	53.50	51.20	28.00	32.90	22.30	20.20	18.30	17.10
Direct Field Cost/Case	13.40	7.50	8.90	6.60	8.80	11.30	7.40	6.60	6.50	6.00
Total Field Cost/Case	18.70	18.00	11.10	9.90	12.60	15.30	9.00	7.70	7.50	7.30

In Sudman's view standards of sampling, processing and control are determined by the nature of decisions to be based on the studies. However, the extra time typically afforded in probability call-back sampling due to longer data collection periods and for the development of analyses strategy and review of pilot material is not always used effectively. If one considers only direct field costs, made up of interviewing costs and supervision, then the cost differential drops to about 2:1 in favour of quotas. The major difference appears to be that of supervisory costs. Not all this difference can be attributed to differences in sampling procedures, more time has to be given to training interviewers in call-back strategy, for example interviewers are expected to check back with supervisors after three attempts. In summary, quality checks and standardised procedures are deemed a necessary overhead of field management.

Sudman also presents important evidence to show that the marginal costs of call-backs is not as large as generally believed. Marginal costs are essentially defined as the one extra hour per unit the average interviewer in a call-back sample must spend to find his/her respondents, as well as additional travel expenses. Table 10 from Sudman shows the marginal travel costs of additional calls. They remain fairly constant, except where the number of cases becomes small. The allocation of travel costs is not straightforward; Sudman distinguishes between travel to and from psu's and travel within psu's. In the former travel costs are allocated equally to completed interviews made on that trip. Not-at-home calls (whether with or without an appointment) are not charged unless there were resulting successful calls. This is justified on the basis that as far as travel time to the psu is concerned additional calls are extra "jam". Travel time within a psu is defined as all time not spent on the interview, and are allocated to all calls made in the segment on that particular trip. As Sudman indicates this type of travel time (doorstep protocol, telephoning, making appointments, polite conversation engaged in to secure cooperation or conclude exchange) may not always be easily discernible from interviewer time sheets (e.g. may get included under interviewing time).

TABLE 10.7:
Average travel cost and marginal cost to complete interview
by number of calls for NORC call-back samples

CALLS REQUIRED TO COMPLETE INTERVIEWS	1			2			3			4		
	AVERAGE		MARG.									
	N	TOTAL COST		N	TOTAL COST		N	TOTAL COST		N	TOTAL COST	
1	792	\$3.23		1,202	\$2.89		7,285	\$.89		3,894	\$1.12	
2	791	4.14	0.91	738	3.50	\$.61	2,562	1.34	\$.45	631	2.55	
3	349	5.30	1.16	480	3.72	0.22	1,187	1.98	0.54	293	3.65	
4	152	6.98	1.68	215	4.43	0.71	661	2.50	0.52	103	4.33	
5	64	8.46	1.48	112	5.24	0.81	351	3.13	0.63	79	4.88	
6	34	8.67	1.21	42	7.06	1.82	176	4.22	1.09	35	6.22	
7 or more	29	9.22	0.55	77	8.13	1.07	219	5.97	0.5	48	9.35	
Total N	2,211			2,866			12,411			5,083		

(taken from Sudman, 1961)

10.3.3:

Durbin and Stuart's experimental study on call-backs and clustering in sample surveys

The study considered was one of a series of experimental studies of survey problems planned by the Survey Research Committee consisting of representatives of the B.B.C's Audience research department, the British Institute of Public Opinion, the British Market Research Bureau, the Government Social Survey, Research Services and the Division of Research Techniques at LSE. The purpose of the paper was threefold: to investigate the performance of call-back procedures, the interviewing costs of different methods of sampling from the Electoral Register and a comparison of different interviewers working under similar conditions. Emphasis will be given to the first objective.

The study is primarily important here because it manifests tangible criteria for design evaluation in the language of bias, variance and cost.

Regarding call-back evaluations attention is confined to cases in which it would be at least possible for a successful call to be made during the field data collection period. Refusals, illnesses and temporarily absent respondents are separated off as categories of non-achievement whereas respondents too busy at the time of call or not-at-home are felt to be attainable by persistent recalling or the making of appointments by the interviewer. Four strategies for dealing with non-response are considered:

- (i) recall up to a specified minimum calls, assumed to be three (Gray and Corlett, 1950).
- (ii) complete recalling on a subsample of unsuccessful calls at the first attempt, similar to double sampling as proposed by Hansen, Hurwitz and Madow (1953).
- (iii) one call only or Politz Simmons (1949) and
- (iv) a replacement quota.

These schemes were evaluated by using a composite index of efficiency containing the principal ingredients of evaluation already discussed, namely cost, bias and variance.

The standard for comparison for these methods and, hence "bias" estimation, is that of "persistent recalling". The index is defined as the reciprocal of the product of mean square error and unit cost.

In addition to the consideration of the impact of non response bias three different types of clustering were considered for six (urban) experimental areas, always with two interviewers from each of three organizations per psu, organised in one of three ways:

- (1) each with a systematic sample of 30 interviews
- (2) each psu with two second stage clusters (polling districts selected with pps) with 15 interviews per polling district and
- (3) as in (b) but with further subdivision of second stage clusters into streets selected pps again to ensure balanced subsamples of 30 interviews each.

The design facilitates both investigation of clustering effects and the comparison of non-sampling variability (differences between interviewers). Partial balance was also achieved by arranging for each organization to use each method in two out of the six areas.

Altogether 32 questionnaire items were compared. Using the summary notation, R_1 , R_3 , R_T , P.S, Q_1 , Q_2 , Q_3 explicit in the extract below. Standard errors in the R_1 , R_3 and P.S columns relate to the differences between those columns and the R_T column, whereas the one for the R_T column relates directly to the attribute considered. For these R_T percentages estimates for each interviewer used in the experiment were calculated to reflect the clustering type.

TABLE 10.8:

**Results for different methods of dealing
with non-response**

QUESTION AND ATTRIBUTE ANALYSED (PERCENT)	RANDOM SAMPLE RESULTS				QUOTAS FOR REPLACEMENT OF NON-RESPONSE AT		
	AT 1ST CALL	AT 3 CALLS	AT ALL CALLS	BY POLITZ- SIMMONS	1 CALL	2 CALLS	3 CALLS
	R ₁	R ₃	R _{T_r}	(P.S.)	Q ₁	Q ₂	Q ₃
2a. Women who are h/wives							
	92.6*	89.3	89.1	91.2	86.6	88.3	88.3
	1.36	0.49	1.40	1.30			
	(258)	(450)	(565)	(247)	(580)	(580)	(581)
3. Of those with jobs who work							
full-time	79.4 ⁺	88.0	88.1	84.3*	88.9	89.4	88.9
	2.37	0.60	1.40	1.38			
	(141)	(483)	(578)	(136)	(606)	(630)	(542)
4. Workers employed in manufacturing trades							
	22.9*	30.3	31.3	21.4 ⁺	22.2	26.6	28.3
	3.41	0.85	2.15	1.68			
	(140)	(482)	(576)	(135)	(600)	(627)	(640)
5. Of full-time workers who work days							
	71.4*	78.8	81.1	74.3 ⁺	80.8	81.2	80.7
	3.27	0.77	1.86	1.91			
	(112)	(425)	(503)	(108)	(536)	(563)	(571)
6. Workers whose time to work is less than 20 mins							
	54.0	54.0	52.3	50.7	61.0	58.1	55.4
	3.67	0.96	2.04	2.20			
	(139)	(462)	(557)	(134)	(561)	(592)	(605)
7. Who went to pictures in last 7 days							
	28.7	29.3	30.3	27.2*	32.4	30.7	29.7
	1.85	0.56	1.51	1.14			
	(373)	(822)	(937)	(354)	(1,074)	(1,078)	(1,081)
8a. Who own a television set							
	30.8	31.1	32.6	33.4	30.3	30.6	30.8
	1.88	0.51	1.57	1.52			
	(374)	(823)	(937)	(355)	(1,074)	(1,079)	(1,082)

(Taken from Durbin and Stuart, (1954)).

* Difference significant at the 5 per cent level

+ Difference significant at the 1 per cent level

Attempting to summarize these findings Durbin and Stuart present the following summary table for the probability samples:

TABLE 10.9:
Resume of information in Table 10.8

	Samples		
	R_1	R_3	P.S
Number of differences significant at: 5 per cent level	6	4	9
Number of differences significant at: 1 per cent level	7	-	4

The R_3 sample (historical standard of the time) has four significant differences out of 32, as against less than two to be expected. The authors indicate that the excess would not be significant even if the different questions were independent, and they are not. However the samples R_1 and P.S provide conclusive evidence for biasedness. For further detail on the results the reader is referred to the original text, however a convenient mechanism for making overall comparisons is provided by the authors by ranking the difference between the R_T result and the six other methods, in the following table:

TABLE 10.10:

Rank analysis of table 10.8

	Method						Total
	R_1	R_3	P.S	Q_1	Q_2	Q_3	
Sum of ranks	146	57.5	134.5	143.5	103	87.5	672

Note: the sign of any difference has been ignored, where there are ties average ranks are allotted. Ranks are low to high.

The table brings out clear disparities. R_3 stands out as the best method, however there is little to choose between R_1 and Q_1 methods. Results on average costs for the three types of clustering are given in tables (10.11, 10.12 and 10.13 below):

TABLE 10.11:
Average cost of interviews on clustering
type 1

Call	No of Interviews at ith Call	No of Interviews in Call up to ith	Average Cost of a Successful ith Call	Average Cost of Interviews Obtained from Calls Up to ith	Standard Error of Estimate a_i
i	n_i	$\sum_{r=1}^i n_r$	a_i	$\sum_{r=1}^i n_r a_r / \sum_{r=1}^i n_r s(a_i)$	
1	109	109	0.187	0.187	0.035
2	108	217	0.199	0.193	0.036
3	30	247	0.304	0.206	0.097
4	14	261	0.268	0.210	0.128
5	3	264	0.753	0.216	0.339
>5	4	268	0.541	0.221	0.240

(Durbin and Stuart - 1954)

TABLE 10.12:**Average cost of interviews on
clustering type 2**

i	n_i	$\sum_{r=1}^i n_r$	a_i	$\sum_{r=1}^i n_r a_r / \sum_{r=1}^i n_r$	$s(a_i)$
1	118	118	0.114	0.114	0.030
2	99	217	0.156	0.133	0.042
3	33	250	0.125	0.132	0.071
4	21	271	0.241	0.141	0.102
5	4	275	0.582	0.147	0.308
>5	10	285	0.334	0.153	0.144

(Durbin and Stuart - 1954)

TABLE 10.13:

Average cost of interviews on clustering
type 3

i	n_i	$\sum n_r$	a_i	$\sum n_r a_r / \sum n_r$	$s(a_i)$	a_i (Excluding Stoke Newington)
i	1	1	1	1		
1	83	83	0.165	0.165	0.039	0.154
2	76	159	0.151	0.158	0.050	0.152
3	42	201	0.184	0.164	0.068	0.199
4	24	225	0.178	0.165	0.099	0.166
5	19	244	0.291	0.175	0.123	0.212
>5	9	253	0.258	0.178	0.107	0.346

(Durbin and Stuart - 1954)

Average costs are obtained by least-squares estimation, where interviewer day is defined as the unit of cost. a_i is the cost of a successful i th call, and n_{ij} the number of interviews at the i th call on the j th interviewer day then estimates of a_i are obtained by minimizing

$$\Sigma (1 - n_{1j} a_1 - \dots - n_{6j} a_6)^2 \quad (10.7)$$

summing over all interviewer days, type of clustering. Suffix 6 denotes calls beyond the fifth. Type 2 turns out to be superior to 1 and 3. For types 1 and 2 average cost increases steadily with call, whereas for 3 it remains fairly constant until the fifth call.

Average costs for type 3 are still less than for type 1 (systematic sampling within psu). The authors suggest that the closer proximity of potential respondents in 3 may have resulted in prior warning of interviewer presence in an area or interviewer fatigue as a result of intensive exploration of a small neighbourhood, resulting in a large number of unproductive calls being recorded (call-back means per 30 attempted interviews for each cluster type are respectively: 56.8, 60.6, 75.9). Of course, the discovery that a clustered sample costs less does not necessarily imply better value for money. Positive within cluster variation will mean higher variance. Thus variances as well as costs are examined by Durbin and Stuart. Using an "accuracy ratio" (or "design effect" where the denominator refers to a two stage unclustered design) for the ratio of a variance for a clustered sample to that of an unclustered sample of the same sample size they show for a sample of m clusters and n individuals within each cluster the ratio will normally be greater than 1 (assuming between cluster variance will be greater than zero). Indeed the accuracy ratios are greater than unity for most of the items for both types of clustering (table 11, text). Also in the majority of cases the effect is greater for type 3 clustering than type 2. Making simplifying assumptions about the presence of other costs Durbin and Stuart portray an impression of the relative efficiencies of the three types of clustering. Assuming first stage units to be ordinarily selected as part of a two stage process they say the accuracy ratio will be 2, such that the "within variance" component is equal to the "between variance" component, and consequently if a clustered sample of accuracy ratio r is introduced at the second stage the overall variance is increased in the ratio $(1+r):2$. For the cost ratio it is assumed that the interviewers time represents about half of the total costs, also if interviewing costs in the clustered sample are about a fraction c of the costs in a type 1 sample then total costs are reduced in the ratio $(1+c):2$. Taking as a measure of efficiency the reciprocal of the product of overall variance and total costs, the ratio of efficiency of the clustered sample to the unclustered sample is:

$$\frac{4}{(1+r)(1+c)} \quad (10.8)$$

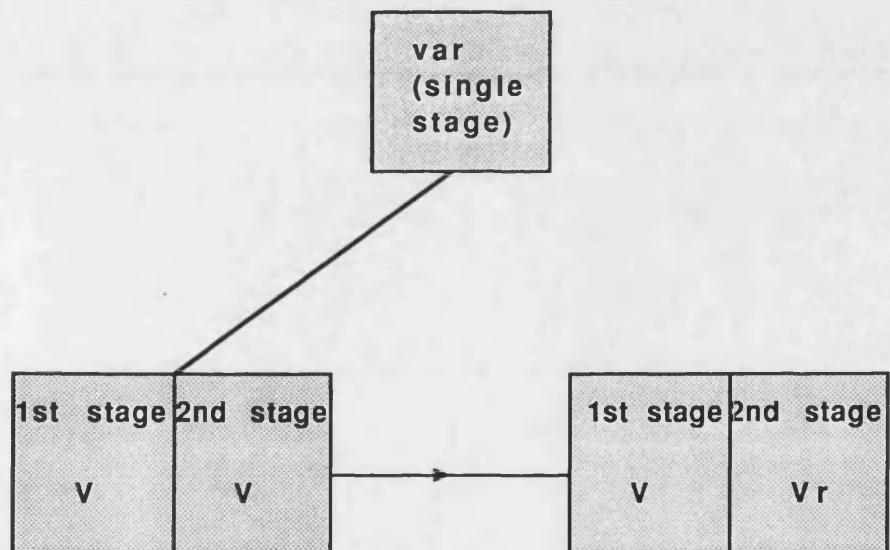
The reader will probably observe that the form of this ratio is similar to the one proposed by Kish, presented in section 10.2.1, where v represents clustered designs and v' unclustered, see (10.5)

The following diagrams in Figure 10.2 may help to make the reasoning more interpretable:

FIGURE 10.2:

**Schema to help interpret the
interplay of costs and variance**

(a) Variance [2 stage type 2 or 3] : variance [two stage type 1]

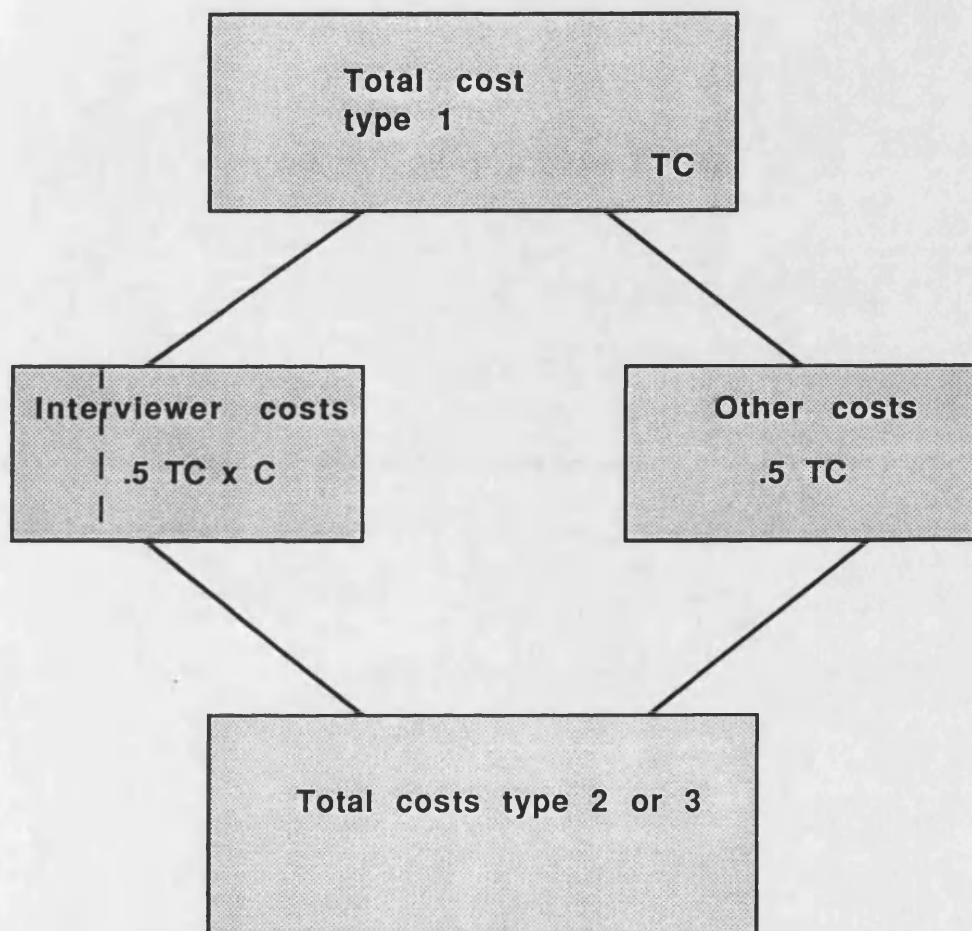


$$\begin{aligned} \text{variance [2 stage type 1]} \\ = 2V \end{aligned}$$

$$\begin{aligned} \text{variance [two stage type 2/3]} \\ = V + Vr \end{aligned}$$

$$\begin{aligned} \text{So variance [type 2/3]} : \text{variance [type 1]} \\ \equiv 1+r : 2 \end{aligned}$$

(B) Costs



Cost type 2 or 3: cost type 1
 $(.5 \text{ TC}) \times C + .5 \text{ TV} : \text{TC}$

1 + C : 2

For samples of type 2 the authors assume C to be around 2/3 thus requiring $r < 1.40$ to result in an improvement in efficiency. For samples of type 3 C is assumed to be about 4/5 suggesting $r < 1.22$ for improvement in efficiency (not very different for the values 1.50 and 1.25 reported, non-interview cost and first stage of variance component assumptions notwithstanding).

Finally, the results on interviewer variability are interesting. Response rates by type of clustering showed significant differences, except for type 1, but reported interviewer differences on items were so small compared with random fluctuations that they could reasonably be ignored. A finding at variance with subsequent investigations of "factual" items (Kish, 1965, O'Muircheartaigh, 1976).

10.4:

The search for optimal design

10.4.1:

Efficient design criteria

In the work of Kish (1965) and Durbin and Stuart (1954) there is a homogeneity of approach towards the development of criteria for design evaluation. The resulting criterion is a product of cost and variance. To use a product, the results may reflect compensating movements in costs and variance, or occasions when both elements behave in a similar way. Taking the reciprocal of this product appears sensible and convenient in that as the product increases, either as a result of increased costs or variance, or indeed both the index diminishes exhibiting low 'efficiency'. The term "efficiency" will be adopted to describe the reciprocal of the product of variance and cost. Thus maintaining efficiency in design would imply an increase (or decrease) in cost should be equivalent to a decrease (or increase) in the variance. Deming (1953, p.135) arrives at the same point by a slightly different route. Firstly he defines the amount of information in an estimate as the reciprocal of its' variance, then "information per unit cost" as $1/(\text{variance} \times \text{cost})$.

To accomodate "total survey error" in any evaluation implies the inclusion of a bias term in the index. This simply means substituting "mean square error" for variance in the product above. Efficiency then could be seen to reflect the "quality of information" in an estimate. An example of the application of this index will be presented in chapter 13. To complete this chapter attention will be given to the work of Kish (1976) to explore how it may be possible to extend the evaluative methodology to multipurpose/item designs.

10.4.2:

Multipurpose allocation

All of the adjustment procedures and framework for evaluation provided by the previous sections and chapter 9 have been developed around a univariate theory. Clearly, out of tune with the multiple objectives of most surveys. Multipurpose surveys are characterised by four main features: the opportunity to measure several respondent characteristics on one visit or contact, the yield of domain or subclass information, repeated observations on the same individual and the yield of a variety of statistics, e.g means, aggregates, differences, correlation coefficients, F-ratios etc., Each of these dimensions multiplies survey objectives, interactions between them further embroider the multiplicity of possible aims. Some of the resulting aims may well conflict. Kish (1976) presents a pragmatic framework to assess such conflict. His work is summarized in the remainder of this section.

Kish unified the approach to problems of optimal allocation by viewing them as distinct components of the same simple expressions for the total variance and cost of the sample statistic y ; expressed in linear forms as follows:

$$\text{var}(y) = V + V_o = \sum V_i^2 / m_i + V_o \quad (10.9)$$

$$\text{cost}(y) = C + C_o = \sum c_i m_i + C_o \quad (10.10)$$

These linear forms are held to occur in multivarious design contexts, including multistage and multiphase designs.

The subscript i denotes the i th component of variance in a design with unit variance V_i^2 for m_i sampling units for that component.

Similarly c_i denotes a unit cost. Components may refer to strata, stages or phases. The V_i^2 and c_i are parameters for which values are assumed or guessed; Kish admits that it might be more realistic to "guess" distributions for these values using a Bayesian approach but abandons the idea in favour of presenting a methodology in reach of survey practitioners. The constants C_0 and V_0 do not affect optimal allocations of the m_i . Other necessary assumptions are listed below:

$V_i^2, \sqrt{c_i}$, V_0 and C_0 non-negative (though in practice V_0 and V_i can be negative). For computing purposes $m_i \geq 2$ and for practical purpose $0 < m_i < M_i$ where M_i denotes the number of population units for the i th component.

Thus the final objective for sample size allocation is to find optimal values of m_i labelled m_i^* where $m_i^* \propto V_i / \sqrt{c_i}$ by minimizing the product:

$$VC = (\sum V_i^2 / m_i) (\sum c_i m_i) \quad (10.11)$$

When either V or C is fixed Kish demonstrates this is the same as finding optimal values for the product $(V + V_0)(C + C_0)$. The product form has obvious connections with the earlier discussion on "efficiency" criteria. Its' form also leads directly to expressions for "loss" functions $1+L$, where L is a relative loss due to departures $k_i \propto m_i^* / m_i$ from the optimal, where, of course, L will be zero.

"Loss" can represent relative increase of variance or cost. It is here the full pragmatism of survey practice and experience evolves. Formally by strict "optimist" standards Kish argues for losses of 2 per cent and 50 per cent on two designs would both be non-optimal, however a "proximist" would usually class a 2 per cent loss with the optimal, to distinguish both from larger losses like 50 per cent. This argument is persuasive especially when we are often faced with design evaluations based on guesses or crude adjustment procedures.

In the general formulation of the loss function Kish defines two parameters, $U_i = V_i \sqrt{C_i} / \sum V_i \sqrt{C_i}$, the relative "sizes" ("efficiencies") of the components and the k_i (expressed above) to reflect relative departures of sample sizes m_i from the optimal allocations. The principal form of the function is obtained by first dividing the product VC by $(\sum V_i \sqrt{C_i})^2$ to compensate for their units of measurement to enable the function to attain the minimal (optimal) value of 1. Finally the function is shown to have the form:

$$1+L = (\sum U_i k_i) (\sum U_i / k_i) \quad (10.12)$$

When $L=0$ all k_i are equal. This is neatly demonstrated with a Lagrangian identity (refer section 5 in the original text).

To illustrate a conflict in allocation two examples taken from Kish's paper provide an excellent application for a stratified design with two strata comparing an instance of minimizing the variance of the overall mean versus that for minimizing the difference between two means.

(a) Consider the variance of the mean $\sum W_i y_i$ for two strata where $W_1 = 0.2$, $W_2 = 0.8$, $S^2_1 = S^2_2 = S^2$ and $c_1 = c_2 = c$. Then $U_i = W_i \propto V_i$, and $U_1 : U_2 = 1:4$. This implies (5.3) that optimal allocation of sample sizes should be in the ratio of stratum sizes W_i , hence $m_2 = 4m_1$. If samples of equal sizes $m_1 = m_2$, are taken, this implies a departure factor of 4; we can use simply $k_1 = 1$ and $k_2 = 4$. The consequent relative loss L would be given by (2.3), 10.12 above, as $1+L = (0.2 \times 1 + 0.8 \times 1/4) (0.2 \times 1 + 0.8 \times 4) = 1.360$.

(b) To illustrate the effect of the U_i on the loss L : suppose now $S_1^2 = 4S_2^2$ and $c_1 = 4c_2$. Since $S_1^2/c_1 = S_2^2/c_2$, optimal allocation is still 1:4. But now $U_i = W_i S_i/c_i / \sum W_i S_i/c_i$, hence $U_1 = U_2 = 0.5$. Therefore the relative loss L from equal sample sizes now would be given by $1+L = (0.5 \times 1+0.5 \times 25) (0.5 \times 1+0 + 0.5 \times 4) = 1.5625$.

In practical terms departures from optimal designs are often unavoidable; true and exact values of cost and/or variance may not be available, convenient sampling fractions may result in departure. For multipurpose objectives there may be different optima, the nature of the sampling frame may provoke departure, the design itself may rest on faulty reasoning or finally departures may result from constraints imposed on the methodology. Table 10.14 below represents diverse frequency distributions for selected population weights, U_i , for both discrete and continuous versions (subscripted d and c respectively). In the discrete versions the relative departure k_i take distinct values from 1 to K , in continuous versions they vary continuously from 1 to K .

For small K the loss L is fairly small and uniform. For $K=2$ to around 5 losses are moderate and fairly similar. Basically below $K=10$ Kish suggests we can make good guesses about L just from the range 1 to K without knowing much about the U_i (provided its' distribution is not dichotomous or U-shaped). Beyond $K=10$ losses increase and diverge.

When dealing with sample results the population weights U_i are no longer appropriate. It then becomes convenient to use sample weights based on sample sizes $u_i = U_i/k_i$. Then the basic product formulation yields various practical expressions (refer 2.4 to 2.7 text) for the loss in terms of the relvariance C_k^2 of the k values with sample weights u_i around their mean $\bar{k} = \sum u_i k_i / \sum u_i = 1/\sum u_i$. Table 10.15 below illustrates six useful examples based on means and variances of convenient finite distributions as presented in Kish (1965, p.262).

TABLE 10.14:

Relative losses (L) for six models of population weights (U_i); for discrete (L_d) and continuous (L_c) weights : for relative departures (k_i) in the range from 1 to K

Models	K	1.3	1.5	2	3	4	5	10	20	50	100	500	1,000
Dichotomous $U(1-U)$													
(0.5) (0.5)		0.017	0.042	0.125	0.333	0.562	0.800	2.025	4.512	12.005	24.50	124.5	249.5
(0.2) (0.8)		0.011	0.027	0.080	0.213	0.360	0.512	1.296	2.888	7.683	15.68	79.7	159.7
(0.1) (0.9)		0.006	0.015	0.045	0.120	0.202	0.288	0.729	1.624	4.322	8.82	44.8	89.8
Rectangular	L_d	0.017*	0.042*	0.125*	0.222	0.302	0.370	0.611	0.889	1.295	1.620	2.403	2.746
$U_i \approx 1/K$	L_c	0.006	0.014	0.040	0.099	0.155	0.207	0.407	0.656	1.036	1.349	2.120	2.461
Linear decrease	L_d	0.017	0.040*	0.111*	0.203	0.283	0.353	0.616	0.940	1.437	1.917	2.879	3.333
$U_i \approx K+1 - k_i$	L_c	0.006	0.014	0.040	0.097	0.153	0.205	0.409	0.680	1.127	1.514	2.507	2.956
Hyperbolic decrease	L_d	0.017*	0.040*	0.111*	0.215	0.312	0.404	0.807	1.466	3.014	5.076	16.802	28.342
$U_i \approx 1/k_i$	L_c	0.006	0.014	0.041	0.103	0.171	0.235	0.528	1.011	2.138	3.621	11.998	19.915
Quadratic decrease	L_d	0.016*	0.036*	0.080*	0.150	0.211	0.264	0.460	0.696	1.048	1.333	2.026	2.331
$U_i \approx 1/k^2_i$	L_c	0.006	0.014	0.040	0.099	0.155	0.207	0.407	0.656	1.036	1.349	2.120	2.461
Linear increase	L_d	0.017*	0.040*	0.111*	0.167	0.200	0.222	0.273	0.302	0.320	0.327	0.330	0.333
$U_i \approx k_i$	L_c	0.006	0.013	0.037	0.088	0.120	0.148	0.22	0.273	0.308	0.320	0.331	0.332

Dichotomous $1+L = 1+U(1-U)(K-1)^2K$
 Discrete $1+L_d = (U_i k_i) (U_i/k_i)$, with $k_i = i = 1, 2, 3, \dots, K$
 Continuous $1+L_c = U k dk. (U/k) dk$, with $1 \leq k \leq K$
 Only two values, 1 and K, were used for L_d for $K = 1.3, 1.5$ and 2

Table 10.15 Losses L for six models of sample weights $u_i = U_i / k_i$; the departures $k_i \leq 1$ from 1 to K represent compensations for undersampling (Kish, 1976)

1	2	3	4	5	6
$\frac{Q}{P} \left(\frac{K-1}{K+Q/P} \right)^2$	$\frac{1+3Q}{3P} \left(\frac{K-1}{K+1+2Q/P} \right)^2$	$\frac{1}{2} \left(\frac{K-1}{K+2} \right)^2$	$\frac{1}{2} \left(\frac{K-1}{K+1} \right)^2$	$\frac{1+2/C}{3} \left(\frac{K-1}{K+1} \right)^2$	$\frac{1}{6} \left(\frac{K-1}{K+1} \right)^2$

Losses (L) can get large for models 1 and 2 when both K and Q are great: $K = 20$ and $Q = 10P$ the L is 4.011 for 1 and 2.219 for 2. For both models losses remain moderate.

The methods for optimization and proximation so far presented are conveniently adapted to multipurpose design; essentially the general form for variance accomodates different purposes by designating purpose by the subscript "g", thus for the g th purpose we have

$\sum_i V_{gi}^2 / m_i$. By assuming that costs are fixed separate loss functions for each variable may be written as

$$1 + L_g = (\sum_i V_{gi}^2 / m_i) / V_{g\min} \quad (10.13)$$

where the denominator denotes the minimal variance attainable for the g th variate. Weights are then assigned to each variate to denote the relative importance of the lost precision of each variate. Kish then considers a total expected loss as a linear function of quadratic loss functions of the variances for a set of fixed m_i . Apparently the resulting expression is a modified version of a function proposed by Dalenius (1957) and related to versions proposed by Yates (1960) and Cochran (1963). Again Lagrangian techniques enable optimal allocations of m_i^* to be established (refer 6.3 and 6.9 text). Choice of weights I_g seem crucial and it would be interesting to witness a practical demonstration of the methodology (letter to Kish 1986) to accompany the theoretical evaluation. Kish regards fixing costs as more practical than trying to fix a set of values for V_g and then to minimise cost. This problem has been demonstrated with "convex programming" (for example see Kokan and Khan, 1967). A review of such work will not be attempted here given the reluctance of most survey practitioners to tackle the ideas already presented (especially the notion of relative importance of purpose). In the absence of a priori knowledge it would seem reasonable that the assumption of "equal" importance would not be misguided and would certainly simplify the expressions involved.

The inclusion of response errors and "bias" estimation has been omitted from the detail presented here so as to assemble a framework closely identified with Kish's own presentation. It would appear feasible to construct a "mean square error" component in the evaluation. It is also proposed to illustrate in Chapter 13 how by using an assumption of "equally important" items a global evaluation of call-back strategies may be attempted.

There appear to be no actual designs based on Kish's work, nor indeed any other formal approaches to multipurpose allocation (Kish, 1986). They are clearly needed. Together with an extension to include nonsampling error such development would present a major enhancement to the evaluations conducted in this thesis.

We have seen in the last two chapters how it is possible to conduct an evaluation for one variable in terms of a product of cost and mean square error. Chapters 11 and 12 describe exactly how a particular retrospective evaluation for a '4 call-backs with appointments' strategy was conducted for the Occupational Mobility Survey. Chapter 11 describes the background to the survey and chapter 12 provides a resume of how the evaluation was operationalized in terms of the methods and ideas reviewed in the previous two chapters. Chapter 13 contains the results of the evaluation, also covered in Wiggins (1988).

Chapter 11: Specific survey conditions for the Occupational Mobility Survey

11.1 Background

11.2 Context of Evaluation

11.1: Background

The evaluation presented in Chapter 13 is based on an Irish occupational mobility study designed by Professors Jackson and Iutaka in association with Professor Hutchinson to investigate the determinants of occupational mobility in Northern Ireland and the Irish Republic. (SSRC Grant HR1430/1)

The target population for study were males aged 18 to 64 years at the time of interview living in Northern Ireland (North) and the Irish Republic (South). The geographical and political divide between North and South provides two domains for evaluation. Table 11.1 below summarizes the design. (for further detail see O'Muircheartaigh and R D Wiggins, 1977).

TABLE 11.1:
Sample design summary

		DOMAIN	
		NORTH	SOUTH
STRATA	URBAN	Belfast constituencies - systematic samples	Dublin constituencies - systematic samples
	RURAL	Two stage zoning	Two stage probability proportional to size

Achieved sample size: North, 2416; Response rate 73%
: South, 2291; ditto 79%

Interviewers were instructed to make up to four calls on any potential respondent. If any call proved to be inconvenient for the respondent an appointment was made, alternatively if someone else in the household could indicate when the respondent might be at home an appointment was made on their behalf. The outcome of every call was carefully recorded and subsequently processed.

As reviewed in Chapter 9, this information facilitates the use of modeling strategies based on call-backs to estimate nonresponse "bias". Combined with appropriate cost assumptions, this provided an opportunity for retrospective evaluation rarely afforded in survey analysis.

11.2:

Context for evaluation

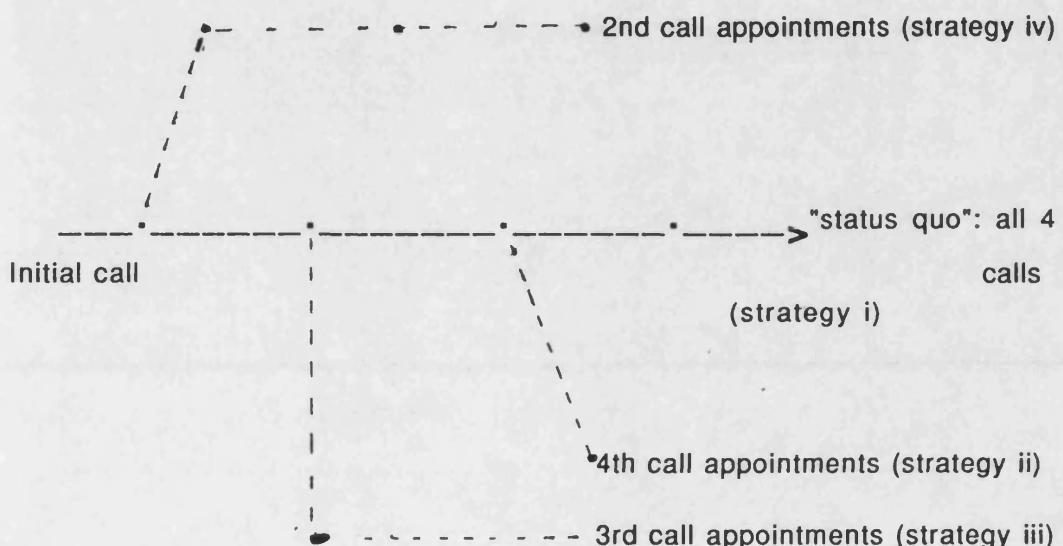
The evaluation is entirely empirical and contains three main ingredients. First, an assessment of the belief that "bias" arises at the initial or early calls in a call-back plan. Secondly, a review of the attraction of appointment calls by examining their relative productivity and bias compared to non-appointment calls beyond the initial call, and, finally an inspection of the data for alternative call-back norms or strategies in terms of a criterion which reflects accuracy and cost.

The call-back norm actually used was the maxim of using up to 4 calls on sampled contacts plus the use of appointments where possible. This maxim will be referred to as the "**status quo**" or strategy (i). Alternative strategies are obviously constrained by this policy. Retaining the notion of an upper limit of 4 attempts to obtain an interview three possible alternative strategies were considered:

- (ii) at the fourth stage of calling only visit appointment calls, strategy labelled "**4th call appointments**".
- (iii) at the third stage of calling only visit appointment calls; as appointments are not always successful at the following stage the outcome of these calls can be traced to a fourth stage of calling, strategy labelled "**3rd call appointments**".
- (iv) at the second stage of calling only visit appointment calls. Again these calls can be traced to a fourth stage of calling, strategy labelled "**2nd call appointments**".

Figure 11.1 below illustrates the four strategies as routes of contacts made which generate responses or successful calls. Note all strategies involve the same initial call, so cannot be considered to be "mutually exclusive". The numbering (ii) through (iv) reflects their degree of closeness to the "status-quo". (i)

Figure 11.1:
Strategies for retrospective evaluation



For any strategy there are four potential stopping points (marked by a dot.) Each strategy can be considered to effect an evaluation 'within' and 'between' routes. For instance, examination of the status quo maxim might reveal the attraction of using one call only, thereby appealing to quota samplers or indeed, the possibility of stopping at earlier calls than the four call norm actually adopted. Alternatively, one can consider the possibility of comparing different strategies at any particular stopping point, e.g. 3 calls versus 3rd call appointments only etc. These possibilities can only be evaluated once the researcher has established an appropriate decision rule to reflect the degree of accuracy and cost associated with each item of information.

Stopping points are denoted in the summaries that follow by numbers in brackets following the strategy title, where (1) denotes the initial stage of calling only, (2) the initial stage plus one recall and so on. Hence for a comparison between strategies "2nd call appointments (2)" would imply that the initial call plus recalls on the appointments made for the second call was considered to be optimal for a particular item.

Altogether six items were considered, see table 11.2 below. Although the author is aware of a need to develop a methodology to synthesize the findings results are largely presented in terms of univariate assessments. An attempt at global evaluation based on Kish (1976) is briefly considered in Chapter 13.

TABLE 11.2:
Items used in the evaluation
Discrete

No. of years to father's full-time education	(var42)
ditto self ditto	(var45)
No. of persons in respondent's household	(var51)

Categorical

Has respondent ever lived outside Ireland	(var38)
ditto been unemployed	(var39)
is respondent a Protestant or of another faith	(var55)

The next chapter summarizes the specific strategy for a retrospective evaluation used in the OMS. The methodology rests on the foundations for modeling non-response bias and survey costs reviewed in Chapter 9 and 10.

Chapter 12:

**A resume of a strategy to evaluate the
impact of non response in a survey with
call backs and appointments**

12.1 An overview

12.2 Selection of cost models

12.3 Use and development of an efficiency criterion

12.1: Overview

The motivation for the evaluation presented in Chapter 13, was stimulated by a concern of Durbin (1953) regarding surveys of human populations:

"... First, could results of equal accuracy have been achieved more cheaply; and, secondly, what should be done to compensate for the non-achieved part of the sample?"

We have seen in Chapter 9 that a major response to the problem of non-achievement in samples has been the use of repeated attempts to obtain responses, be it call-backs, rediallings or reminder mailings. In interview surveys call-back strategies become part of the strategy to control non-response and part of the problem itself. For example, the decision to set a limit on the number of calls and encourage the use of appointments may be interpreted differently by different interviewers. The planning and timing of call-back routes may ultimately affect a respondent's readiness and willingness to cooperate. Survey costs and the quality of information obtained finally bear witness to such variation in operational styles.

Chapter 13 focuses on a specific call-back plus appointments policy as used in the OMS. Emphasis is given to methodologies, notably those of Bartholomew (1961) and Drew and Fuller (1980), to refine estimates of the survey mean in a way that actually takes account of call-back policy. As actual cost information was not made available, cost models are developed in the light of those reviewed in Chapter 10. These models express cost outcome at each stage of (re) calling to enable evaluations in terms of the efficiency criterion defined in Chapter 10. Obviously unless there is a clear relationship between the number of attempts to obtain a response and the value of a survey item the methodologies are unwise. In the absence of much empirical evidence to the contrary this assumption seems reasonable.

As well as providing a framework for retrospective evaluation of call-back plus appointments strategy, Chapter 13 provides an opportunity to explore the notion that most bias occurs in the early stages of a call-back policy and review the relative attraction of making appointments whenever a respondent is too busy at the time of call or "not-at-home".

12.2:

Selection of cost models

Without some estimate of survey costs it becomes difficult to compare alternative strategies of collecting survey data. The cost models presented below condition on the actual design used and only consider alternative call-back strategies identified within the design (see chapter 11). Concurring with earlier work by Hansen, Hurwitz and Madow (1953), Kish (1965) and Sukhatme and Sukhatme (1970) the important cost component in the evaluation is always "average variable cost". However, unlike Kish, costs of sampling and locating contacts are assumed to be part of the overall fixed costs, irrespective of call-back policy. Differential costs associated with travel to and from primary sampling units are also assumed to be part of the fixed cost component. Expressing cost formulations as an outcome at each stage of calling enable "efficiency" indices to be calculated. It is also straightforward to identify the relative efficiency of appointment versus non-appointment calls within any particular stage.

Following Kish (1965) "average variable cost" is defined as

$$\frac{\sum n_i c_i + k}{\sum s_i} \quad (12.1)$$

where n_i is the number of contacts at each stage of calling
 s_i the number of successful calls
 c_i the cost per successful contact

and k the direct interviewing cost, assumed constant throughout (set to 4 arbitrarily).

Different interpretations of the costs of field work practice provide different estimates of c_i . Essentially, c_i represents the travel cost component of average variable costs for each wave. Three cost models are considered. In model I, $c_i = 1$; this is equivalent to using the number of calls to be made at each stage as a measure of cost. This model can only measure costs directly if the mean cost is the same for all calls.

In model II, $c_i = \sqrt{n_1/n_i}$, a term similar to that proposed in equation 10.6, where n_1 acts as a proxy for the size of an area to be covered by all interviewers, and costs are inversely related to the number of contacts to be visited at any stage. In model III, where $c_i = n_1/n_i$ results in the same cost for each set of contacts. In this way, model II represents a compromise between I and II.

Cost model I is likely to underestimate reality in that it assumes a constant cost per contact irrespective of whenever an interviewer is visiting 1 or 1001 contacts (an assumption also made by Thomsen and Siring, 1983). Cost model III is at the other extreme, with model II as a sound compromise. The overall formulation is similar to the framework suggested by Kish (1965). It is limited in comparison to slightly more sophisticated formulations by Hansen, Hurwitz and Madow (1953) and Sukhatme and Sukhatme (1970) in that no allowance is made for travel between primary sampling units or travel from home or office to a p.s.u. Considering this component to be part of the fixed costs may be permissible in urban areas, but less so for rural ones. Kish argues that overemphasis on travel cost ... "can mislead us to overestimate the cost increase for later calls", especially if the fixed cost per interview is high relative to travel cost. He goes on to suggest that the first call cost should ... "include the entire cost of selecting sample cases, finding them, and identifying them.." This would imply placing first (initial) calls at even greater cost disadvantage than is the case presented in Chapter 13.

Clearly the cost models presented here can be criticized for not coming close enough to a complex reality. It was felt until more is known about the efforts interviews make to obtain a response in terms of planning routes and use of appointments it would be preferable to utilize models that are straightforward and easy to interpret.

12.3:

Use and development of an efficiency criterion

To reflect the level of accuracy implicit in a survey estimate we need a global measure which should ideally mirror all sources of survey error. A useful criterion to capture the operation of such sources, is naturally, as discussed in 10.4 "mean square error".

In the empirical evaluation that follows variance reflects sampling error alone. Components of variance due to non-sampling error (e.g. interviewer variability) are not catered for in the sense that the specific survey design does not permit their estimation. As is often the case "bias" is alluded by the absence of a true value for all items. The temptation might be to simply speak of "precision" of any survey estimate as reflected by the variance rather than accuracy. However for the evaluation a "conditional bias" can be obtained which conditions on the actual results of the survey. Thus at any potential stopping point in a call-back strategy the resultant estimate of a survey mean can be compared with the mean actually obtained after four calls. Thus the bias at any particular stage or stopping point of a call-back strategy is defined as

mean obtained for stage of —	actual survey mean (12.2)
call	under 4 call maxim
(a potential stopping point)	i.e., the status quo

In a sense this bias reflects the non-response bias that would have been obtained if call-backs were terminated at a potential stopping point. Resulting estimates of mean square error obtained by this procedure are labelled 'RDW' in subsequent illustrations. In terms of the actual number of responses obtained stopping call-backs at any earlier point than the status quo policy would strictly result in a reduced achieved sample size.

In the evaluation alternative strategies are considered so as to achieve the same number of respondents as in the original survey and also one where 5000 completed interviews were achieved. If any alternative were adopted this would imply inflating the initial contacts made as well as altering the call-back procedure. (bias estimates remain unaffected by changes in sample size, therefore evaluation implies only an adjustment in the variance estimate for different final sample sizes).

"Relative bias" is obtained by considering the absolute value of the difference in 12.2 divided by the actual survey mean after four calls to examine the extent of bias present in early calls and a comparison of type of call (appointment versus non-appointment) within each stage of recalling. Alternative bias estimates are obtained by refining the actual survey mean to take account of some of the methodologies reviewed in Chapter 9. In particular by assuming the the mean obtained at the last (i.e fourth) call is equivalent to the mean likely to be obtained for any outstanding/potential respondents enables a reweighted survey mean to be calculated. This procedure is based on a suggestion by Bartholomew (1961) to allow for "not at home" bias in sample surveys and so will be labelled "BMEW" in subsequent illustrations. The other procedure used applies the maximum likelihood methodology of Drew and Fuller (1981) outlined in 9.4.3 and will, therefore, be labelled "MLE". Applying Drew and Fuller the sample data is first arranged into age call categories with estimates of potential interviews outstanding to comply with the original demands of the model. Parameter estimates are obtained by a non-standard application of the maximum likelihood program, MLP (1985). (for further assistance refer to Dr G Ross, Rothamsted). Revised estimates of the proportion of the sample falling into the various age categories are then used to reweight the survey estimate. Details of the data used and the resulting estimates are contained in Wiggins (1988), (see table 2). Combining estimates of bias and variance at each stopping point provides a measure of approximate "mean square error".

Cost estimates resulting from the use of any of the three models described in 12.2 can be combined with estimates of "mean square error" at every stopping point to reflect on the

relative merits of any of the call back strategies identified in chapter 11. As described in 10.4 the reciprocal of the product of "cost and mean square error" is used to determine "the efficiency" of a potential stopping point in any call-back plan. Given that cost and accuracy can be "traded off" it is quite possible that greater benefits in accuracy may accrue at an increase in costs. This was considered unrealistic in terms of the retrospective nature of the evaluation. Thus any gain efficiency that implied increased costs was ruled out. Some alternatives deemed 'efficient' may be 'proximal' in that they represent decreases in cost but a marginal increase in mean square error.

Chapter 13 is structured in a way to permit a careful examination of all of the features of the "call-back plus appointments" procedure described in the preceding sections. Initially an examination of early calls and type of call (appointment versus non-appointment) will be presented in terms of productivity and relative bias defined using the original survey mean. Then, in terms of the efficiency criterion defined above, the evaluation focuses on a retrospective examination of the four alternative call back strategies (outlined in Figure 11.1) for each of three different procedures for estimating bias at any stage of (re) calling, for each of the three cost models, under two different expected achieved sample sizes (original and 5000).

All evaluations refer separately to each domain in the sample design (North and South) for each of the six survey items described in section 11.4.

Chapter 13:

A global evaluation of non-response in terms of mean square error and estimated survey costs

Contents

Introduction

13.1: Findings

13.2: Discussion

Introduction:

The importance of the evaluation in this chapter is that it provides a rare opportunity to witness a global evaluation of a data collection strategy. Information about costs and relative biases due to non-achievement of the sample at various stages in a call-back strategy is presented to demonstrate a retrospective framework for survey appraisal. Clearly, the methodologies used are open to question, enhancement and improvement. It is intended that chapters 9 and 10 will have stimulated or provided the potential for any refinement. The evaluation is presented in the spirit that only by the accumulation of such empirical information will sensible economic appraisal of survey data become familiar in the survey literature.

The chapter begins with an assessment of relative bias for the "four call norm plus appointments" for the six items selected for appraisal. The appraisal then develops an evaluation strategy in terms of the efficiency criterion and cost models presented in earlier chapters.

13.1: Findings

Table 13.1 below shows the relative bias of successive calls for six items under the original call-back plan conditioning on the survey mean after four calls as a criterion for assessment. In the North the relative bias is never more than 2.5% within the value obtained by the fourth call; typically any bias present at the second calls (for 4 of the six items) demonstrates a steady decline over successive stages. Generally, by the third stage of calling substantial gains in bias reduction have been attained. In the South relative bias tends to be higher at the initial stage (up to 12% within the value obtained by the fourth call), however with two minor exceptions the decline over successive calls is generally steady. Investing in up to three calls appears to be valuable.

Table 13.1:
Relative bias of successive calls for six items over four calls

		var42	var45	var51	var38	var39	var55
Initial call		.0108	.0007	.0108	.0000	.0201	.0218
up to 2 calls		.0046	.0015	.0211	.0008	.2229	.0249
up to 3 calls		.0001	.0005	.0066	.0008	.0057	.0093
up to 4 calls		.0000	.0000	.0000	.0000	.0000	.0000
South							
Initial call		.0376	.0339	.0370	.1234	.0746	.0500
up to 2 calls		.0236	.0166	.0051	.0723	.0658	.0500
up to 3 calls		.0190	.0016	.0004	.0170	.0044	.1000
up to 4 calls		.0000	.0000	.0000	.0000	.0000	.0000

Beyond the initial call each recall is classified according to whether or not an appointment was made at the previous call. Table 13.2 shows how appointment calls were always at least twice as successful than non-appointment calls.

Table 13.2:
Relative success rates for "appointment" versus "non-appointment" calls

	<u>1st call</u>	<u>2nd call</u>	<u>3rd call</u>	<u>4th call</u>
NORTH	-	4.91	3.80	2.51
SOUTH	-	2.27	2.35	2.61

We are now in a position to examine the extent to which appointments realised their obvious productivity in terms of relative bias. Table 13.3 summarizes the outcome. In the North there is no overwhelming evidence to suggest appointment calls provoke relatively less bias than non-appointment calls overall, though clearly researchers are not going to lose by encouraging appointment strategies. (note relative bias is expressed in absolute terms). For most items, particularly the dichotomous ones, the divergence between appointment and non-appointment calls seems to be of greatest magnitude. In the South appointments overall seem to be singularly worse than non-appointment calls (var39 notwithstanding) and this discrepancy seems to be maintained throughout all stages of recalling. Again large amounts of relative bias are observed for dichotomous items at the final stage of recall.

If anything results for the North appear to conform more closely with the idea of careful timing and planning of 2nd and 3rd call appointments over non-appointments as suggested by Bartholomew, whereas the reverse would appear to be the case in the South.

Table 13.3:

**Relative bias for appointment versus
non-appointment calls during recalls
beyond the initial call**

<u>2nd Call</u>		<u>3rd Call</u>		<u>4th Call</u>		<u>Overall</u>		<u>Item No</u>
<u>appt.</u>	<u>non-appt.</u>	<u>appt.</u>	<u>non-appt.</u>	<u>appt.</u>	<u>non-appt.</u>	<u>appt.</u>	<u>non-appt.</u>	
North								
.0058	.0178	.0104	.0303	.0735	.0067	.0000	.0000	var42
.0033	.0020	.0025	.0055	.0152	.0056	.0033	.0031	var45
.0704	.0923	.0338	.0162	.4575	.0863	.0267	.0240	var51
.0583	.2671	.0113	.2100	.5625	.1775	.0479	.0408	var38
.0074	.0842	.0095	.1648	.4327	.0367	.0034	.0017	var39
.0533	.1181	.414	.0822	.2656	.0964	.0170	.0267	var55
South								
.0104	.0033	.0841	.0208	.0610	.0889	.0262	.0077	var42
.0250	.0005	.0884	.0288	.0057	.0202	.0369	.0111	var45
.0711	.0010	.0698	.0348	.0421	.0012	.0607	.0185	var51
.0613	.0243	.2417	.0664	.2515	.1396	.1136	.0357	var38
.0487	.1316	.0008	.1833	.2259	.1189	.0000	.0008	var39
.2850	.0008	.0750	.3800	.2650	.1925	.2450	.0625	var55

It may be instructive next to consider the three different estimation procedures (labelled RDW, MLE and BMEW) used to define "bias" due to non-achievement at various stages in each possible retrospective call-back policy. Table 13.4 below summarizes the results.

Table 13.4
Estimates for overall means and proportions
used in evaluation

var	NORTH			SOUTH		
	<u>RDW</u>	<u>BMEW</u>	<u>MLE</u>	<u>RDW</u>	<u>BMEW</u>	<u>MLE</u>
42	8.209	8.209	8.205	8.677	8.720	8.688
45	10.343	10.346	10.342	10.411	10.421	10.429
51	4.270	4.255	4.271	4.667	4.664	4.669
38	0.240	0.239	0.241	0.235	0.237	0.236
39	0.349	0.350	0.349	0.228	0.229	0.228
55	0.642	0.648	0.647	0.040	0.039	0.040

The range of the results for the different estimates are encouraging. All estimates appear quite similar. For variables 42,45 and 51 estimates both North and South are all within 0.5% above or below the conventional/conditional survey estimate (RDW). For variables 38,39 and 55 the majority of estimates are within 2/3% of the RDW estimate, though for variable 55 in the South the divergence is larger (around 7%). There is no systematic trend in the rank order of the estimates obtained. The RDW estimate tends either to be the lowest or middle ranked value; whereas the BMEW estimate tends to be the highest ranked value on the majority of estimates (7 out of 12).

Before proceeding with a presentation of the results for all items for each estimation procedure under each cost model it is felt that the reader will gain a better appreciation of the steps involved in the evaluation by illustrating the methodology for a single item (no. of years of full time education var45) in a single domain (North) using the RDW estimate for bias under strategy (iv), 2nd call appointments.

Table 13.5 provides full information about the components of mean square error.

TABLE 13.5:

Illustration of evaluation for a single variable, no. of years of full-time education (var45) strategy (iv), 2nd call appointments North domain. RDW estimation method

<u>Stopping Points</u>	<u>Mean</u>	<u>Rel.bias Ratio</u>	<u>Variance</u>	<u>Bias</u>
(1) 1st call	10.336	0.0007	2.8949E-03	0.1301
(2) 2nd call appts	10.322	0.0021	2.6548E-03	0.4270
(3) 3rd call	10.316	0.0026	2.6767E-03	0.5219
(4) 4th Call	10.315	0.0027	2.6407E-03	0.5449

Overall parameter estimate = 10.343

Variance denominator = 2416

Note: The bias ratio is "the absolute bias \div an appropriate standard error"

The relative biases are small at every stage, however beyond the first stage they show a steady increase. Given reasonably constant variances across all stopping stages the bias ratios tell a similar story to the relative biases. Thus based on both of these criteria alone continuing to call beyond the first stage would appear unwise! The evaluation proceeds with consideration of average variable cost at each stage as well as bias and variance. It is therefore necessary to introduce a set of cost assumptions; the one used in this illustration is captured in cost model III, using 12.1 , $c_i = n_1/n_i$ and $k = 4$. With data presented in Wiggins (1988) an average variable cost component can be calculated. Consider the second stage of calling, under this strategy an initial call plus a recall on any appointments made at that first contact. In particular,

$n_1 = 4433$ with successful calls, $s_1 = 691$
 and $n_2 = 1033$ with successful calls, $s_2 = 821$

Thus the average variable cost,

$$= \left\{ \frac{4433 + 1033 \times \left(\frac{4433}{1033} \right)}{1033} \right\} / \frac{(691 + 821) + 4}{4}$$

$$= 9.8638$$

with a mean square error (from table 13.5) of
 $(.022)^2 + 2.6548E-03$.

Finally, resulting in an efficiency score (the reciprocal of the product of a.v.c and m.s.e) of 32.2992. (also see table 13.6 below). Indeed the efficiency scores in table 13.6 uphold the findings in table 13.5.

The highest efficiency score is the one obtained at the first stage, though despite the powerful cost assumptions captured in model III continuing to the second stage of calling almost looks attractive (indices 32.6143, 32.2992 respectively). Figure 13.1 completes a graphical illustration for this within variable evaluation.

TABLE 13.6:

**Efficiency measures for a single item: number of years of full-time education strategy (iv) :
 2nd call appointments North domain, RDW estimation, cost model III**

Stage

Efficiency index	(1)	(2)	(3)	(4)
Sample Base, 2416	32.6143	32.2992	23.9090	19.6867

Figure 13.1:

Efficiency scores plot for strategy (iv) 2nd call appointments for cost modell III item 45, North domain (sample base 2416). Based on table 13.6)

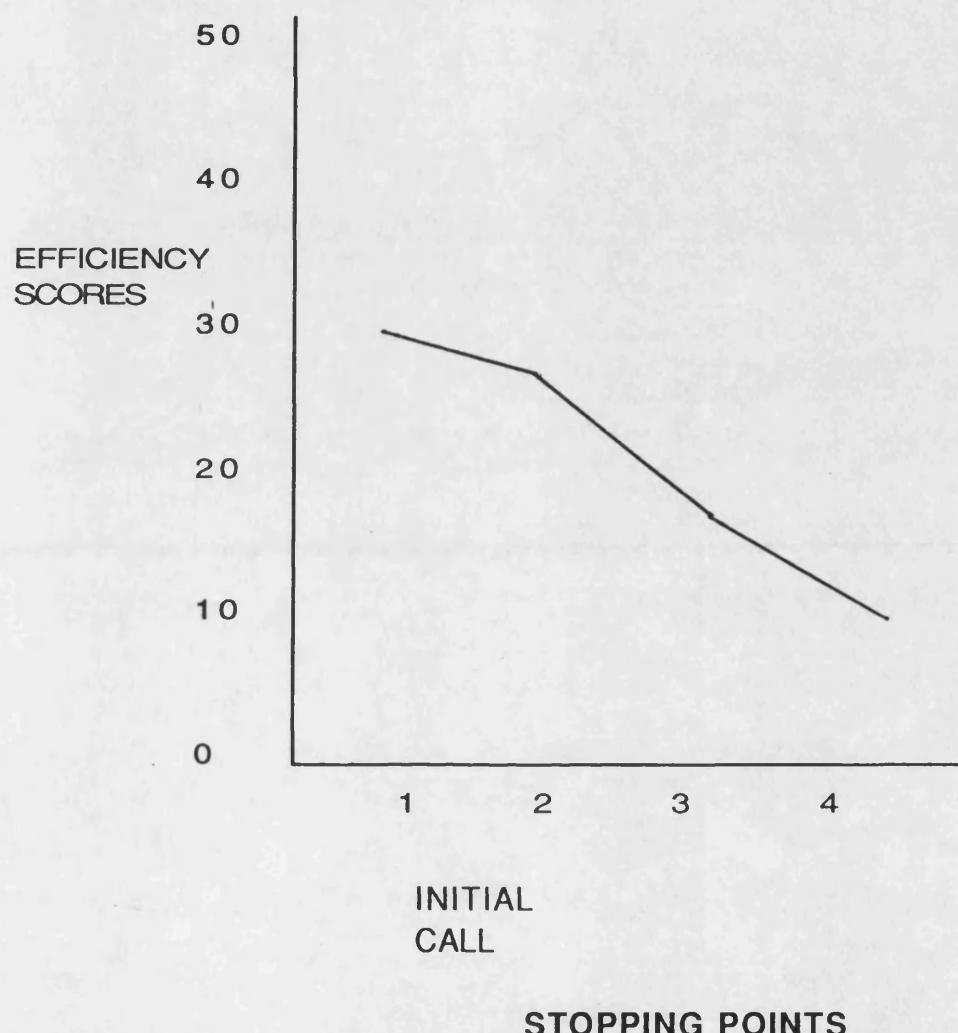


Table 13.7 completes an evaluation for variable 45 in terms of efficiency scores and relative efficiency measures for achieved samples of 2416 and 5000 respectively. Relative efficiencies are defined in two ways:

(i) by conditioning on the previous stopping stage, so that any ratio >1 would indicate a marginal increase in efficiency and an indication to proceed to the next stage of calling,

or

(ii) by conditioning on the final stopping stage, so that ratios with a value >1 indicate that there is little gain in continuing with the policy of making calls to the fourth/final stage.

For either sample bases or efficiency criteria there seems no reason to qualify the earlier decision to continue calling beyond the first stage.

TABLE 13.7:

**Efficiency measures for a single variable
no. of years of full-time education (var45)
strategy (iv) 2nd call appointments North
domain RDW estimation/cost model III**

Efficiency index		Relative efficiencies					
<u>Sample Base</u>		<u>Condition on prior stopping point</u>		<u>Condition on final stopping point</u>			
2416	5000	2416	5000	2416	5000		
32.6143	66.3159	1.0000	1.0000	1.6567	2.0262	(1)	
32.2993	57.3812	0.9903	0.8653	1.6567	1.7532	(2)	
23.9090	40.2629	0.7402	0.7017	1.2145	1.2302	(3)	
19.6867	32.7290	0.8234	0.8129	1.0000	1.0000	(4)	

It is now possible to extend this evaluation to include all possible cost models and call back strategies for all items. Tables similar to table 13.7 were produced for all six items and form the basis for the graphical summaries included in Wiggins (1988) for all cost models and strategies. In particular, consideration of cost model I and II for variable 45 in the North first indicate a different conclusion to the one reached above; under both models efficiency criteria would suggest calling to the second stage (see table 13.8 below).

TABLE 13.8:
Efficiency indices for variable 45 under cost models I and II under strategy (iv)
North, RDW

COST MODEL I		COST MODEL II		
<u>Sample</u>	<u>Base</u>	<u>Sample</u>	<u>Base</u>	
2416	5000	2416	5000	
32.6143	66.3159	32.6143	66.3159	(1)
41.8371	74.3254	38.1677	67.8067	(2)
39.1171	65.8735	34.0421	57.3271	(3)
39.0901	64.9869	33.0970	55.0233	(4)

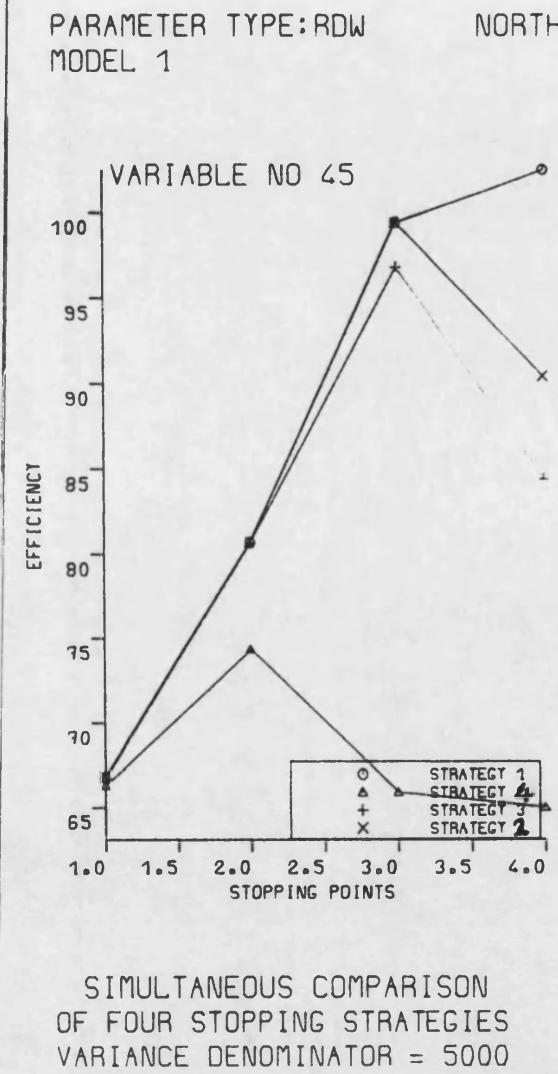
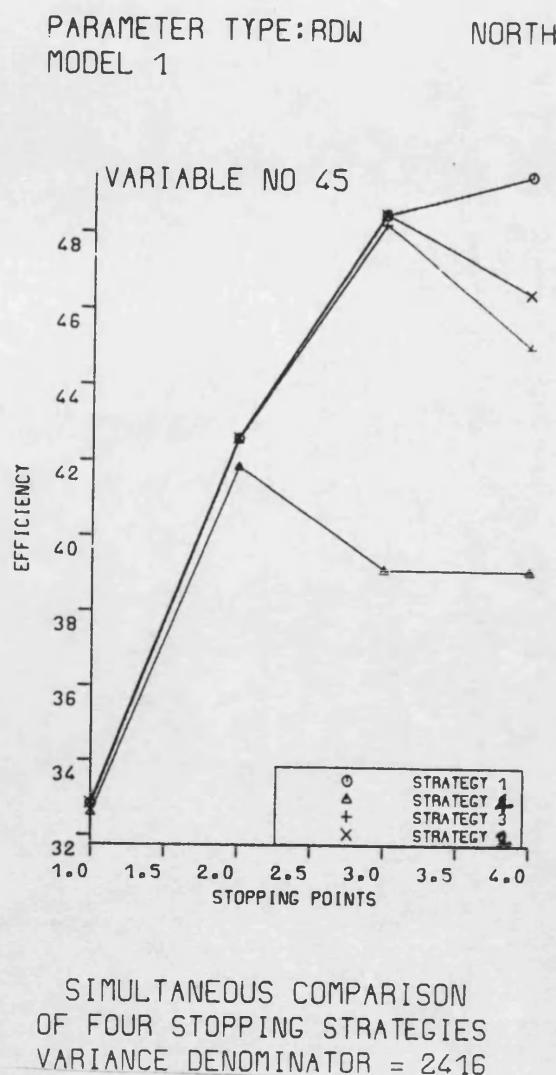
Figure 13.2 extends the evaluation to a simultaneous comparison of all strategies. Efficiency criteria under model III for a sample base of 2416 indicate that it would be desirable to continue calling to the second stage of calling for all strategies except strategy (iv). This is because under other strategies going to the second stage means both a lower bias ratio (0.3082 cp 0.4270) and lower average variable costs (9.0605 cp 9.8638). Thereafter all strategies do badly. Changing the sample bases to 5000 alters the conclusion slightly for strategy (i) (and at this stage it is equivalent to strategy (ii)) where the efficiency score suggests stage three as a stopping point. By increasing the sample base the relative reduction in the mean square error makes the move between stage 2 and 3 attractive. By definition average variable costs remain the same irrespective of the sample based used (stage 2 cp 3 gives 9.0605 cp 9.9159). However, under sample base 2416 the relative reduction in the mean square error is 7% between these stages but 15% under a sample base of 5000 (base 2416: mse: 2.9518E-03 cp 2.7265E-03: base 5000: mse: 1.5586E-03 cp 1.3304E-03).

Under cost models I and II the conclusion for different strategies changes; in all strategies except strategy (i) there are grounds to continue calling until the third stage. Whilst this is also true for strategy (i) under cost model II it is not the case under cost model I, where completing all four stages of calling (the status quo) is confirmed when contrasted to all other strategies.

Clearly it is possible to continue this mode of analysis to include consideration of the other procedures for estimating the overall mean. Introducing BMEW and MLE procedures only provide minor qualifications to the above conclusions. Under a sample base of 2416 MLE leads to exactly the same conclusions as under RDW; the BMEW estimate provokes a change to three calls compared to RDW's 2 calls under cost model III. Increasing the sample base to 5000 again replicates the RDW conclusions for the MLE estimate. On this occasion the BMEW provokes the only instance of change, namely support for the status quo under cost model II. A detailed summary of these findings can be found in Wiggins, 1988.

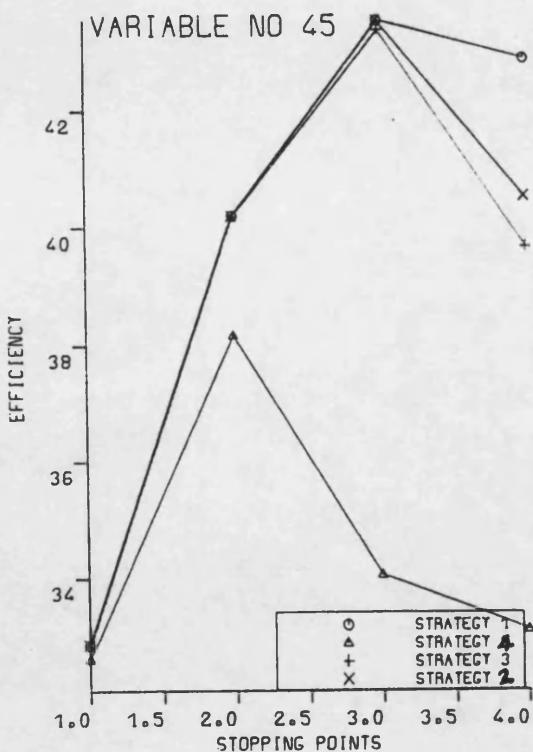
Figure 13.2:

An illustration of simultaneous comparison of four stopping strategies across three cost models in the Northern domain



PARAMETER TYPE: RDW
MODEL 2

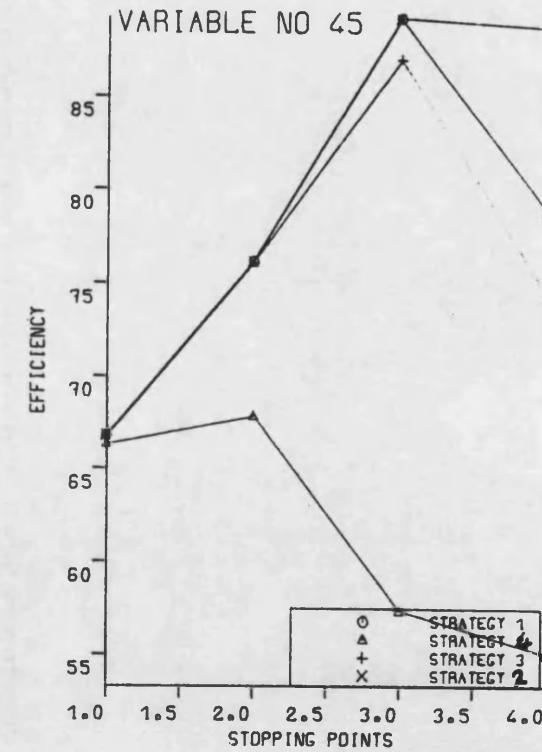
NORTH



SIMULTANEOUS COMPARISON
OF FOUR STOPPING STRATEGIES
VARIANCE DENOMINATOR = 2416

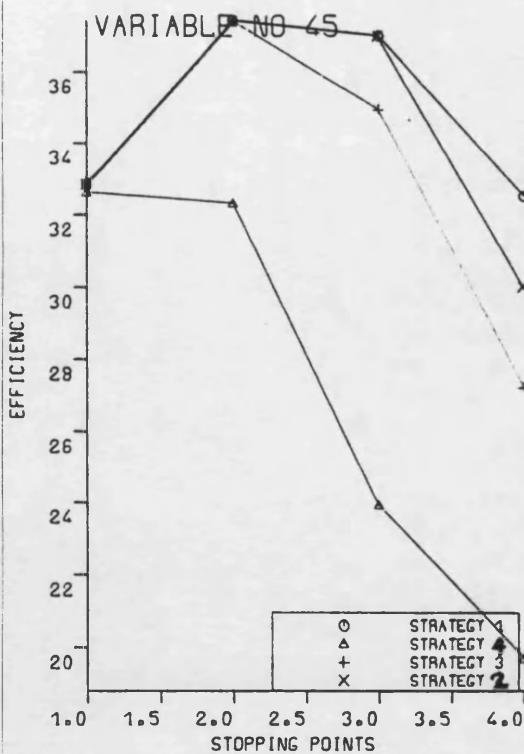
PARAMETER TYPE: RDW
MODEL 2

NORTH



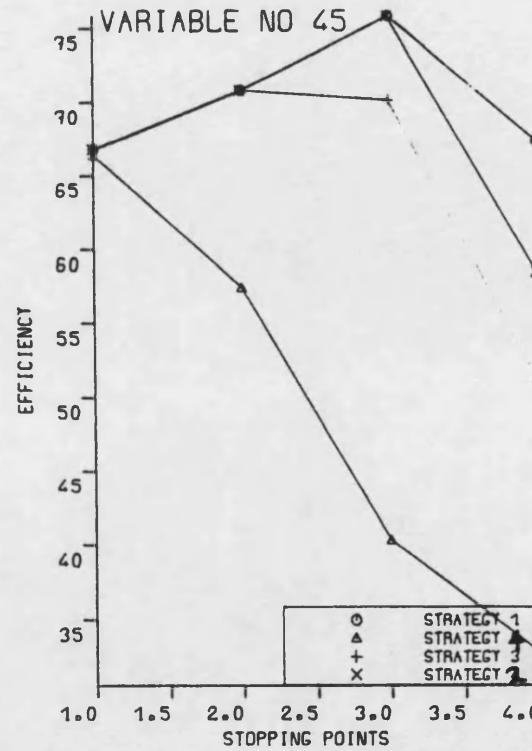
SIMULTANEOUS COMPARISON
OF FOUR STOPPING STRATEGIES
VARIANCE DENOMINATOR = 5000

PARAMETER TYPE: RDW
NORTH
MODEL 3



SIMULTANEOUS COMPARISON
OF FOUR STOPPING STRATEGIES
VARIANCE DENOMINATOR = 2416

PARAMETER TYPE: RDW
NORTH
MODEL 3



SIMULTANEOUS COMPARISON
OF FOUR STOPPING STRATEGIES
VARIANCE DENOMINATOR = 5000

Thus, attempting a summary for the North in terms of cost models, model I would provide strong comfort for the conventional norm of all four calls. Under cost models II and III wherever there is any divergence from the status quo it tends to be towards the strategy of making up to three calls (both appointment and non-appointment). For certain items there is even a suggestion for two call strategies.

Again in the South cost model I assumptions support the status quo - however, the results are somewhat different for different items. On this occasion items 42,45 and 38 almost always indicate "status quo" calling irrespective of the cost model. Where item 51 diverges under cost model II and III it is generally towards a three calls strategy. Again item 39 throws up 2 or 3 call strategies under cost models II and III item 55 appears erratic varying between 2 and all four call routes. Closer inspection of the graphical presentation for this item reveal a curious divergence between the third call appointments route and the status quo at the third stage which appears to smooth out by the final stage. This "hiccup" seems to be provoked by a rather large bias term at this stage for the status quo which diminishes by the final stage of calling i.e the third call appointments route has a consistently smaller bias at the third stage for all estimation procedures (-.002 vs .004 for both RDW and MLE; and -.001 vs -.003 for BMEW); this divergence is also reflected in the relative efficiency scores. Rarely does increasing the sample base alter these conclusions.

13.2:

Discussion

In the Northern domain initial calls appear to believe the general belief that most of the bias arises at that stage, though relative bias is within very acceptable (proximal) items limits across all stages of calling for most of the items considered. However results for early calls in the South do appear to accord more closely with Bartholomew's view that most of the bias arises in the initial stages. Certainly in both domains we witness substantial reductions in bias by the third stage of calling.

Appointment productivity looks attractive; particularly, in the North. Though as regards relative bias the relative attraction is less appealing, especially during the fourth stage of calling. The separation between appointment and non-appointment calls at this stage remains somewhat curious. Perhaps individuals who agree to appointments after continued efforts to gain cooperation are more like "refusers" than the busy "not at homers" obtained by chance at the final stage.

Doubling the sample size in either domain does not appear to substantially alter the findings based on the original achieved sample sizes. This is largely due to the fact that the relative reduction in mean square error has not been substantial enough. Given that bias remains unchanged by increases in sample size this implies that the bias : variance relationship reflected in the bias ratio has remained fairly constant.

The major practical implication of the evaluation represents a challenge to fieldwork management, practice and questionnaire design. Regards fieldwork, taking up an alternative call-back strategy, eg. only make up to 2 calls, would imply inflating the initial number of sampled contacts so as to ensure the same achieved sample size as in the original survey. In the North this would mean increasing the number of contacts to be sampled by a factor of 1.379. The suggestion assumes no change in the quality of the information as reflected by its' mean square error. Table 13.9 is used to illustrate the process.

TABLE 13.9:

An illustration of going to two calls in the North so as to maintain the original achieved sample of 2416

(a) Original productivity

	1st stage	2nd stage	Totals
Contacts	4433	2494	6927
Successful calls	691	1061	1752

On the basis of (a) we expect 9552 contacts will generate 2416 successful calls; an inflation factor of 1.379.

(b) Reappraised allocation of contacts, constant success rates

	1st stage	2nd stage	Totals
Contacts	6113	3439	9552
Successful calls	953	1463	2416

Note: Average variable costs remain constant under (b)
e.g. under cost model III

$$\begin{aligned}
 \text{a.v.c} &= \left\{ 6113 + 3439 \times \frac{6113}{3439} \right\} / 2416 + 4 \\
 &= \left\{ 4433 + 1061 \times \frac{4433}{1061} \right\} / 1752 + 4 \\
 &= 9.0605 \text{ as under (a)}
 \end{aligned}$$

The inflation of initial contacts in this manner does have serious repercussions for fieldwork management. One might be for agencies to use more interviewers over a shorter fieldwork period, desirable both in terms of interviewer effect (O'Muircheartaigh, 1977) and the avoidance of any deterioration in performance over time (Cannell et al., 1970). Though there is no overwhelming assurance that the quality of the data would remain intact. Another interesting possibility is the idea of "double sampling" (Hansen, Hurwitz and Madow, 1953) which in theory could accommodate the possibility that different call-back strategies are suggested for different items. There would be two types of questionnaire, one which consisted of items for which limited call-backs were felt necessary and another consisting of items requiring a more intensive call-back policy. Clearly, preservation of the coherence of the survey questionnaire would present practical problems which might render such a suggestion unworkable. Furthermore all of the evaluations condition on the survey design actually used. An interesting departure for extending the evaluation might be to consider clustered versus unclustered designs (refer to 10.3.3), as attempted by Durbin and Stuart (1951). Such an opportunity might be afforded by running evaluations on an urban-rural dichotomy within each domain. Unfortunately, this proved practically unattainable due to the ad hoc nature of the survey.

The bias estimation procedures considered may well be an issue. There are different estimates simply because there is no single methodology for refining survey estimates which has universal appeal. All procedures conditioned on the survey itself-all give seemingly reasonable parameter values, with RDW and MLE estimates in close accord (table 13.4). As regards application of the Drew and Fuller approach understanding is still developmental. An extension of the modelling procedure to allow for differential response patterns between those responding to appointment calls versus non-appointment calls was attempted but not utilised. It was considered that whilst the model appeared sensible in terms of interpretation its parameters were very poorly estimated (Ross, 1983).

The other component of 'efficiency' is, of course, cost. Hopefully, the cost models have an intuitive appeal. Ultimately,

all of the models depend on assumptions regarding travel. This may be limited in comparison to slightly more sophisticated formulations by Hansen, Hurwitz and Madow (1953) and Sukhatme and Sukhatme (1970) in that no allowance for travel is made between primary sampling units (p.s.u) or travel from home or office to a p.s.u. This may have been a reasonable omission for tightly clustered urban areas, but less so for the rural allocations. Additionally, Kish argues that an overemphasis on travel cost 'can mislead us to overestimate the cost increase for later calls', especially if the fixed cost per interview (always 4 units in our case) is high relative to the travel cost. He argues that the first call should include the entire cost of selecting cases, finding them and identifying them. This would imply placing first/initial calls at a greater cost disadvantage than is presented here.

In practice, it is likely that cost model I is likely to underestimate by assuming a constant cost per contact irrespective of whether interviewers make one or a thousand and one contacts (where $x_j = 1$). Thomsen and Siring (1983) used a similar assumption in a recent application. Cost model III is at the other extreme, where costs are set by the initial spread of first calls ($x_j = n_1, n_j$) and model II is regarded as a sensible compromise ($x_j = \sqrt{n_1, n_j}$).

Simply using the efficiency index without proper regard to behaviour of its' constituent parts may be unwise. An improvement in efficiency could imply four outcomes: 'more cost, less m.s.e', 'less cost, same m.s.e', 'less cost, larger m.s.e', 'less cost, less m.s.e'. Unfortunately, none of the results included the latter outcome! Outcomes embracing the first option were ruled out on the assumption of fixed budgeting. For illustration these have been underlined in table 3 of the Appendix in Wiggins (1988). Ideally, the practitioner would need to decide on acceptable (proximal) limits for any 'trade off' between cost and mean square error.

A final reservation is that all evaluations are conducted separately for the six individual items. Typically such practice whilst in accord with sampling theory, contradicts the multipurpose nature of survey design in chapter 10 where Kish's framework for global multi-item evaluation was reviewed.

A simple idea gleaned from his work might be to begin by averaging efficiency scores across items for each strategy on the basis that each item has equal worth in terms of information. For instance under strategy (i), the original four call maxim, in the North for cost model II we witness a steady improvement in efficiency as the number of calls increase (table 13.10 below).

TABLE 13.10:

Average efficiency score for six items under strategy (I) in the North for cost model II

Stopping point	Score
1	9.4352
2	14.1969
3	23.7718
4	26.6441

Formal approaches to review multipurpose allocation are both needed and possible (Kish, 1986) if survey researchers are to satisfy themselves that data is always collected in a way which maximises all of the information regarding the survey process (not simply item response but call-back outcome, strategy and management as well). Despite the reservations expressed here and the obvious room for further refinements there appears to be no apparent reason why such evaluations should not become routine and commonplace in survey reporting. It is the view of the author that it is not enough to simply accept fieldwork norms on call-backs and appointments without attempting an evaluation of outcomes.

Part Two: Conclusion

The principal source of nonsampling error considered in this part of the thesis has been non-response bias. For a survey mean we have seen that this bias is the product of two components, the actual level of reported non-response and the difference between the non-respondent and respondent means. In interviewer surveys the reported level of non-response is the final outcome of fieldwork endeavour. Data analysis typically proceeds on the basis of the responding elements. There is no alternative. The nature and sources of non-response were reviewed at the beginning of chapter 9. There are important lessons to be drawn from the work of Hansen, Hurwitz and Madow (1953), Kish (1965), Proctor (1977), Sandstrom (1977), Van Westerhoven (1978), Steeh (1981), Thomsen and Siring (1983), Lievesley (1986) and Morton-Williams (1986) in order to develop or improve fieldwork strategies in future surveys. Their work and ideas will play an important role in the final chapter by 'bridging' the evaluations presented in both parts of the thesis.

It is seldom possible to determine the exact nature of the second component of non-response bias, namely the difference between the respondents and non-respondents on an particular characteristic, without external validating information or non-respondent follow-ups. Most survey reports merely report the level of non-response, thus implicitly assuming this component is zero. It is possible to do more. By conditioning on the survey itself estimates of the population mean can be obtained by varying assumptions about the relationship between the survey variable and the amount of effort required to obtain a response, typically the number of calls or dials. This facilitates a 'what if' evaluation, e.g. what would be the (conditional) bias of our estimate if we had continued calling up until the i th call rather than the call-back norm actually used? In this manner a conditional estimate of bias can be obtained for any potential stopping point in the call back scheme. Three methods of estimating bias were used; each method provided an estimate of 'truth' which was compared with the mean actually obtained at a particular stopping stage to produce a bias term, the methods for obtaining estimates of 'truth' were:

- i) simply using the actual survey mean obtained after four calls (labelled RDW)
- ii) reweighting the survey mean to take account of the number of outstanding calls; the mean for the last call was assumed to be equal to the mean of the outstanding calls. (labelled BMEW)
- iii) using a maximum likelihood procedure based on the relationship between the number of calls and the age structure of the population for them response variable. (An approach based on Drew and Fuller (1981), labelled MLE)

An estimate of bias can then be combined with estimates of other sources of survey error, notably sampling error, to establish a global criterion to capture the level of accuracy associated with a survey estimate. This term was entitled 'approximate mean square error' to denote its dependence on necessary assumptions to produce estimates of bias and omissions due to failure to obtain estimates of other sources of nonsampling error. Most investigators would agree that the maximisation of accuracy (the inverse of mean square error) is a worthy aim, but one which cannot make practical sense unless the aim is 'framed' by budgetary considerations, e.g. by using mean square error as an outcome can researchers realise more accurate, or least no less accurate, information at no additional cost? For these reasons the evaluation illustrated in chapter 13 combines mean square error and cost to define an index of efficiency for a survey variable to enable a closer reflection of reality.

The necessity for repeated call backs to ensure the fruits of rigorous sample design and implementation is often seen as a deterrent to probability sampling. An interesting by-product of the evaluation is that first calls could be regarded as analogous to quota sample calls. Bartholomew (1961) suggests that most bias arises at the initial call. Results for the Northern domain in the OMS study belie this belief, which might be good news for quota samplers. However, this is not the case in the South.

For both domains substantial reductions in bias were witnessed by the third stage of calling. It is only by providing more empirical evidence of this kind that better call back planning and budgeting can be facilitated.

If there happens to be no obvious relationship between the value of a survey variable and the number of attempts required to obtain a response, then there is no need to invest in repeated calling. Only when more evaluations are accumulated will it be possible to reflect on such an assumption. However, bias estimation procedures may remain an issue. Other methods might also be tried to estimate the 'truth'. Two groups of procedures, namely those which make adjustments to the call back scheme on the basis of *a priori* assumptions (see 9.4.1) and those which take no account of the number of attempts required to obtain a response (see 9.4.2) were reviewed but not subsequently illustrated. Practically, methods under the first group (Politz and Simmons (1949, 1950) and Bartholomew (1961) influenced the choice of methods actually used in chapter 13. Both of these strategies might be thought of as 'reduced call back' schemes resulting in one or two calls only. Their success depends largely on interviewers being able to gain accurate information about a potential respondent's likelihood of being at home. These methods are referred to as 'indirect' as they depend on *a priori* assumptions rather than the 'post hoc' adjustments made under the other methods that derive their origins from the more general procedures applicable to incomplete data e.g. Rubin (1974, 1976), Chapman (1976), Platek and Gray (1979), Rubin and Little (1987).

The decision to implement a four calls plus appointments scheme in the OMS study had to be taken as a 'given' prior to evaluation. This determined the retrospective character of the evaluation. The illustration centred on alternative uses of call-back schemes in order to question conventional wisdom about how many calls are necessary to secure data quality for a given resource outlay. What did actually influence fieldwork management in the OMS study was the attention Bartholomew (1961) drew to the value of appointments. Indeed, an interviewer's use of appointments looked extremely attractive in terms of producing a successful call (the relative success rates of appointment versus non-appointment calls was between 2.4 and 4.9). However, the attraction was less appealing in

terms of relative bias. An extension of the maximum likelihood modelling procedure to allow for differential response patterns between those responding to appointment calls versus those responding to non-appointment calls was considered. Whilst the model seemed intuitively sensible its' parameters were poorly estimated (Ross, 1983) and, therefore not presented.

As indicated earlier, the other component of the efficiency index was cost, the absence of actual cost information drove the review of potential cost models presented in chapter 10. Hopefully, the three cost models actually considered (labelled I to III) have some intuitive appeal. All models condition on the number of interviews actually obtained and therefore assume that overheads or fixed costs would be the same under any alternative scheme. This implies that 'average variable cost' becomes the appropriate component. The models make no allowance for travel between primary sampling units or travel from home or office to a p.s.u. This may have been reasonable for urban areas but less so for rural areas. More sophisticated formulations, for instance those advocated by Hansen, Hurwitz and Madow (1953) and Sukhatme 1970) could be tried. Further, Kish (1965) argues that the first call should include the entire cost of sampling cases, finding them and identifying them, thus putting the first call at a greater disadvantage than subsequent calls.

In the illustration, cost model I was felt to lead to underestimates of the true cost as it assumes a constant cost per contact irrespective of whether interviewers make one or one thousand contacts. Thomsen and Siring use a similar assumption in a recent application (1983). Cost model III is at the other extreme, by fixing costs of calling at any stage to the cost of the initial call. Cost model II is regarded as a sensible compromise between I and III.

Simply using the efficiency index (the reciprocal of mean square error times average variable cost) without paying close attention to the relative changes in its constituent parts may be unwise. An improvement in efficiency can imply: 'more cost, less m.s.e', 'less cost, more m.s.e', 'less cost, same m.s.e' or 'less cost, less m.s.e'. Any outcome involving the first option was ruled out, and unfortunately none of the results involved options of the last type. This implies alternatives wherever suggested will be 'proximal' rather than 'optimal'.

For the Northern domain, results under cost model I always support the strategies involving all four calls. Under the other cost models wherever there is a divergence it is towards making three calls, whether for all three calls or third call appointments only at the third stage. In some cases there is support for a two calls only strategy. In the Southern domain, cost model I underwrites the four call strategy for each of the six items considered. Under the other cost models, the results diverge for different items. Half of the items evaluated confirm use of the status quo, all four calls, whatever cost model is considered. The remaining items indicate that a call back plan based on something between two or four calls might be acceptable.

Different estimation procedures for bias only provoke slight modifications to these conclusions. Similarly, altering the assumption about the desired sample base rarely alters the findings. The main implication for fieldwork agencies is turn the focus of attention away from levels of achieved response to the idea of inflating the number of initial contacts (illustrated in chapter 13) for a given call back plan. This is a result of shifting the emphasis away from the impact of the level of nonresponse to the nature of the relationship between the respondents and the nonrespondents. Living with low response rates may well make researchers feel uncomfortable. The other implication is that if any agency did shift its; call back plan it may be wise to use larger fieldforces to cope with the inflation of initial contacts. This could be advantageous in terms of interviewer variability (O'Muircheartaigh, 1976) but difficult in terms of management.

Apart from any weaknesses in the apparent realism of the costing models the major problem with the evaluation is that it is conducted in terms of single item appraisals. This contradicts with the multipurpose nature of survey design. Refining the methodology in the context of a proximal/optimal framework as suggested by Kish (1976) would be a great enhancement. A useful starting point might be to simply average efficiency scores across item sets. In this way entirely different call back strategies might be supported, or different plans for different geographical areas.

Whatever the outcome it is hoped that the illustration presented in chapter 13 will stimulate refinements of this sort

The impact of nonresponse should not be viewed in isolation. Its' presence is due in part to the behaviour of the interviewers themselves. We have seen in the first part of this thesis how basic outcomes regarding an individual interviewer's performance, like reported level of response, can be incorporated into modelling strategies to explore the interrelationships between survey variables. In this way, related aspects of the data collection process, interviewer effect and nonresponse, can be reflected in data analysis. We have also seen in the second aprt of the thesis how these components of nonsampling error can be included as constituents of mean square error to facilitate an evaluation of the quality of data for a given outlay of resource. The final chapter summarises the findings for both aspects of the methodological investigation of nonsampling error in the context of developing a global framework for the evaluation of data collected by means of survey interviews.

Chapter 14: Summary And Speculation

Contents:

- 14.1 Interrelated aspects of the methodological investigation of non sampling error**
- 14.2 Summary of main findings and recommendations**
- 14.3 Future work**

14.1: Interrelated aspects of the methodological investigation of nonsampling errors

The success of any survey operation can be measured by the degree of closeness between the conception of the survey, which will include the definition of the target population and ideal parameters, and its' inferential population, denoted by a set of conceptual units generated by the data selection, collection and processing stages for a given set of survey conditions. Differences arise between these populations from the operation of sampling and nonsampling errors. Sampling errors arise because only a subset of the population is measured in the survey. Nonsampling errors arise from measurements with poorly formulated questions, misunderstood questions or as a result of unintended interviewer effects (response errors), from failure to measure some number of the sample (non response error), and from excluding certain members of the population (non coverage). In chapter 1 we saw that like sampling errors, some nonsampling errors have no effect on averages across repetitions of the survey (defining a single survey as a trial). They simply increase the variation in the values obtained (as measured by response variance). By adjusting the survey design, it is possible to estimate some sources of response errors, notably the effect that different interviewers have on the responses. In the first part of the thesis, particularly in chapter 7, we discuss the variance of measures arising from the differential impact of interviewers for three separate experiments. Although interviewer training attempts to standardise the behaviour of interviewers, the interaction between interviewer and respondent, tends to result in variations across interviewers. Most survey analysts assume that these errors balance out over the sample and that the unbiased nature of survey estimates will be maintained. One of the main tasks of the thesis has been to challenge that view, and demonstrate that for fairly modest variation across interviewers large inflation of sample variance estimates is witnessed across a wide range of survey items and summary scores. Other nonsampling errors create biases in all sample estimates, e.g. by the extent to which nonrespondents differ from respondents, failure to measure all of the members of the sample create bias. Typically, these errors cannot be estimated from the sample itself, without external data. Though, data analytic techniques do make it possible under certain

simplifying assumptions (see review in chapter 9) to obtain 'bias estimates'. In the second part of the thesis three procedures for bias estimation conditioning on the survey itself are illustrated in the context of a retrospective evaluation of a call back strategy. These evaluations demonstrate how is it possible to include estimates of bias in a measure of 'total survey error', in particular mean square error ($bias^2 + variance$), to explore the achieved level of accuracy (the reciprocal of mean square error) for particular potential stopping points in the call back stages of data collection. To properly 'contextualise' the evaluation the assessment criterion needs to be modified to include the costs of data collection. Thus the purpose of the investigation becomes 'could results of equal accuracy have been achieved more cheaply, and, secondly, what should be done to compensate for the non-achieved part of the sample?' (Durbin, 1954). Compensations for the non-achieved part of the sample are obtained via bias estimation procedures to facilitate an operational definition of 'accuracy' and combined with estimates of cost to realise a measure of 'efficiency' (the reciprocal of the product of mean square error and cost) for a single survey item. Throughout the evaluation in part two it is assumed that estimates of sampling variance can be obtained routinely for probability samples. (For complex samples see Wolter, 1985). The two parts of the thesis begin to have common ground, in the sense that the second part establishes a framework for assessing the quality and precision of survey estimates collected by means of interviewers. The design modification and resulting estimation context reviewed and illustrated in part one permit components of variance that describe 'interviewer effect' to be included in our definition of 'total survey error' along side other components of nonsampling and sampling error. Provide resources permit there is no reason why other sources of nonsampling error, e.g. coder variability, could not be included as well.

The two parts of the thesis have been separated for practical and historical reasons. The potential for more 'holistic' assessments of survey data, is clearly charted. Its' realisation is yet to be demonstrated in the literature, perhaps indicating the tensions between the day to day pressures of survey design and implementation with the interests of methodologists. The empirical illustrations covered in parts one and two taken together fall somewhere between a specially designed study for the investigation of nonsampling error and routine survey in

which mere 'lipservice' is paid to the potential for distortion in the responses due to nonsampling error. The results in chapters 7 and 13 emphasise the interrelationships between decisions made prior to, and during, the implementation of data collection so as to present the reader with evidence to challenge implicit assumptions and norms in survey practice.

Incorporating any methodological investigation as part of a survey strategy will always have an associated cost. Apart from anything else this stands as a major deterrent to improving our understanding of the impact of nonsampling error on the quality of survey data. But, those costs must set against the costs of actually carrying out any survey in the first instance. Without such information the debate between survey researcher and methodologist cannot be properly conducted. Unfortunately, survey cost reporting is all too rare in the survey literature. This thesis is no exception in that the evaluations conducted in part two depend on cost models rather than actual costs. It is hoped that work by Sudman (1961) and Groves and Kahn (1979) will remind researchers of the need to remedy this shortfall in global evaluation. To this juncture the thesis demonstrates a global framework for the evaluation of the quality and precision of survey estimates under a given outlay of resources. The other major contribution is the routine inclusion of the presence of interviewers in the modelling of relationships between substantive variables. In this way, interviewers become part of the explanatory process itself. O'Muircheartaigh and Wiggins (1981) began an important contribution to this philosophy. Subsequently the availability of variance component software (Longford, 1986a, 1988a) made it possible not only to include the presence of interviewers in any modelling, but, to also permit the simultaneous inclusion of interviewer characteristics in the modelling as a way of investigating the sources of any interviewer effect. This latter feature provides an appropriate link between the impact of interviewer differences and non response bias. The introduction to chapter 9 provides evidence to question the way most investigators treat non response. Typically the practice is simply to report its' level. No information is provided about non respondents. Thomsen and Siring's work (1983) helps breakthrough conventional approaches which tend to separate out non response from other sources of non sampling error. They consider non response as dependent on a combination of controllable variables, e.g. the selection and training of

interviewers, the call back plan, and indirectly controllable factors, such as interviewer experience, motivation and the respondent characteristics, in a way which naturally combines the process of obtaining responses with the impact of the interviewers. The description of interviewer behaviour is obviously limited by what is measurable. Only a small number of interviewer characteristics were included in the modelling illustrations covered in chapter 8. For the ANS study these were: average number of calls per workload, response rate per interviewer, interviewer gender, and years of interviewing experience. In the PHS study the characteristics included were: average number of calls per workload, interviewer age, gender, supervisor rating of experience and an attitude score reflecting the perceived differences between the disabled and nondisabled. The exciting aspect of the modelling is that the influence of interviewer characteristics can be combined with the substantive analysis in a way that was never possible before. Research on the importance of the 'doorstep introduction' (Morton Williams and Young, 1986) and the 'situational' nature of nonresponse (Lievesley, 1986) provide a 'wealth' of good ideas as to what could be routinely included in the evaluation of call back strategies and studies of sources of interviewer effect.

14.2: A summary of the main findings and recommendations

The interrelated nature of decisions taken at all stages of the survey process not only affect the quality of survey estimates but the relationships between the survey variables as well. Any investigation of non sampling error should reflect the interlocking nature of these considerations. The summary of the main findings of the thesis that follows will be presented under three headings:

- i) the estimation context/implications of design modification
- ii) empirical results, following the implications of i), and,
- iii) the modelling context

References to findings and chapters that support these areas are made in the order in which they might support a global evaluation of the quality of a survey, rather than as constituents of three separate 'adhoc' investigations included in parts one and two.

Taking each area in turn:

i) The estimation context:

the empirical investigation of interviewer variability is only possible with some modification to the survey design so as to include an element of 'randomisation'. Chapter 2 presents a classification experimental designs used to study interview variability. Experiments vary between those whose express purposes is the study of interviewer effect (Gales and Kendall, 1957) and those where interviewer variability studies form only a subsample of the main inquiry (e.g. the ANS and PHS studies used in chapters 7 and 8). The degree of randomisation attainable will typically be a product of researcher commitment and resource availability. The author is in agreement with Collins (1980) that a minimum design requirements for the investigation of interviewer effect in terms of a form of interpenetrating design would not generally be difficult or unduly expensive to achieve. At lease for part of the sample of a survey design. There are also consequences arising from further sophistication in design modification that add further credence to this view. No matter how carefully investigators plan their experimental design, interviewers, by the very nature of interviewer variability, will rarely complete equal size workloads. This renders any design 'unbalanced'. The more factors that are included in the design (e.g. location of interviewer, supervisor, agency) the more complex the resulting inference. Chapter 3 carefully reviews all of the underlying assumptions which stem from the 'design context' for unbalanced data. Three possible scenarios are considered : one and two way classifications and interpenetrating designs. Rarely was the design sophistication exceeded by this coverage in any of the studies of interviewer variability reviewed in chapter 2. Once design modification is secured the analysis of interviewer variability conventionally proceeds by the appraisal of omnibus F-ratios or measures that describe the proportion of variance attributes to the interviewers themselves (Kish, 1962).

All traditional analysis of variance procedures assume that the response variable is quantitative. Seldom will this be the case in survey analysis. For binary items, like those in the FLP scale, with endorsements between 0.2 and 0.8 (Cox, 1970), and summary counts or percentage scores (like the GHQ and FLP scale scores) there is reasonable comfort in using analysis of variance as an indicator of the sensitivity of any effect. For polytomized items alternative appraisals might be constructed using methodologies suggested by Gales and Kendall (1957), Everitt (1977) and Payne (1977). These are briefly demonstrated in chapter 7 for the ANS study. These cautionary comments obviously carry over to any multivariate appraisal (outlined in chapter 4 and illustrated in chapter 7). However, there is a major responsibility on the part of the investigator to carefully consider the level of measurement assumed by his/her appraisal. The main disadvantage in dropping the quantitative assumptions is, of course, that global evaluations based on the use of mean square error are no longer feasible. If quantitative measurement is assumed, serious issues still remain for the analyst. Beyond the routine inspection of F-ratios or roh-values, there is the issue of the inference populations. Do the findings apply solely to the interviewers employed on the study itself, or some wider population of interviewers? i.e. is 'interviewer effect' to be conceptualised as 'fixed' or 'random'? The author prefers fixed effects assumptions on the grounds that most interviewers are specially recruited or persuaded to participate in interviewer variability experiments rather than randomly selected from some hypothetically large pool of willing participants. Even under a 'fixed effects' linear model, individual interviewer effects are not simply obtained for unbalanced data. Effects are not 'estimable' unless 'restrictions' (a logical ingredient of the model specification) or 'constraints' (included in the model specification solely for the purpose of obtaining a solution) are defined. Often these refinements have some intuitive appeal, but sometimes they remain 'data dependent' (as for unbalanced nested designs) and obscure realistic interpretation. In the case where a two way classification is adopted there are 16 possible F-ratios to consider when evaluating the impact of any interviewer effect for unbalanced data. Considerations, such as these must be carefully considered by the investigator before embarking on any study of interviewer effect. The increasing complexity of the interpretation possible generated by deepening design sophistication must surely act as a sobering deterrent to investigators and add further to the idea of using

one way classification or interpenetrating designs for at least part of the sample. These design modifications could be accompanied by much greater use of planned comparisons or contrasts between interviewers as a way of exploring the influence of interviewer characteristics, e.g. experience, gender. Another attractive alternative might be the use of cell means models. These formulations (Searle, 1987) circumvent the problems of 'estimability' and/or the inclusion of appropriate 'restrictions' or 'constraints' in the model specification. However, the onus to develop useful hypotheses in terms of the cell means themselves rests solely with the investigator. The temptation is to relocate hypotheses for these models in relation to conventional analysis of variance. The author is unaware of any demonstrable application of this approach (Searle, 1986).

Extending analysis of variance approaches to include random effects lays the foundation for variance component modelling. The generalisability of the approach lies in defining all linear models to be 'mixed', in the sense that all models have at least one 'fixed' effect, namely the general mean. The history of the development of variance component modelling leading up to the availability of appropriate software is presented in chapter 3. Essentially, traditional analysis of variance models used for the investigation of interviewer effects, like that underlying the one way classification, where

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (14.1)$$

such that μ is a 'fixed' effect and α_i is a 'random' interviewer effect can be extended to permit the simultaneous fitting of other survey variables and interviewer characteristics, such that

$$Y_{ij} = \mu + \beta_1 X_{1j} + \beta_2 X_{2j} + v w_i + (u_i + \varepsilon_{ij}) \quad (14.2)$$

where $\alpha_i = v_0 + v w_i + u_i$

w_i demonstrates an interviewer characteristic

u_i demonstrates an error term associated with modelling
and v_o is subsumed under μ

where interviewer response rates, attitudes etc. might be represented by terms like w_i . Not only is it possible to identify components of interviewer variability in this way, but, to 'tap' the source of this variability as well. The extent to which estimates of 'fixed' effects may vary for individual interviewers can also be explored by allowing these effects to enter the random part of the model. Findings for this approach are summarised under the next heading.

The other component of non sampling error considered in part two is, of course, non response error. The examples presented in chapter 9 (Cochran, 1953), Thomsen and Siring, 1977 and Locker et al (1981) provide evidence to question the implicit assumptions that respondents and nonrespondents are the same in some sense. This motivated a detailed consideration of various procedures for the estimation of 'bias' for survey estimates. All of the procedures considered condition on the survey itself and so the resulting estimates are properly considered as 'conditional'. These estimates are obtained by making varying assumptions about the relationship between the survey variables and the amount of effort required to obtain a response, measured by the number of calls made, as well as type of call (appointment versus non appointment). Three methods are illustrated in chapter 13 to obtain bias estimates at each stage in the callback scheme. The first method conditions on the actual survey mean obtained after all four calls (labelled RDW), the second reweights the number of outstanding potential interviews after four calls by the fourth call mean (labelled BMEW), and the third method uses a maximum likelihood procedure based on work by Drew and Fuller (1981). All of the methods depend expressly on there being a relationship between the value of the survey variable and the number of calls needed to obtain a response. They can be contrasted to other procedures reviewed in chapter 9. These procedures either depend on 'a priori' assumptions about interviewer's ability to obtain information about a respondent's availability (Politz and

Simmons, 1951 or Bartholomew, 1961) or methods which owe their origins to more general procedures developed for handling incomplete data, not simply that arising from failure to obtain a response, but misrecorded or missing data as well as partial nonresponse (e.g. see Rubin and Little, 1987). As a specific 'four calls and appointments' strategy was inherited 'post hoc' it was not possible to implement methods belonging to the first group, though they did influence the rationale underlying the choice of the BMEW method. Procedures under the latter category await trial. They would be especially important if no relationship between the value of the survey variable and the number of calls necessary to obtain a response was deemed to hold. Two important 'offshoots' of the evaluation presented in chapter 13 were to explore the extent to which it is felt that most 'bias' arise at the initial call (Bartholomew, 1961) and the relative gains (if any) of encouraging the use of appointments.

ii) Empirical findings for the estimation context:

for convenience and logical relationship to parts one and two this subsection is divided to reflect results for the interviewer variability studies and the callback evaluation.

a) interviewer variability:

for univariate assessments in the ANS study many of the attitudinal items have variance estimates that would be inflated by a factor of between 1.6 and 3.6 when the differential impact of interviewers is taken into account. For the PHS study, in both years, well over 15 percent of the items has associated interviewer variance inflation factors of between 1.6 and 5.2. Results also include variance estimates for roh-values. The findings for the PHS study were particularly interesting because the experiment was replicated over two separate years. Despite similarities in the average value of roh for each wave (around 0.3) individual items exhibited considerable variability between the waves. This would suggest that a 'once and for all' categorisation of an item as being sensitive to interviewer effect might be misleading. The finding supports the pragmatism implicit in allowing routine regular interviewer variability appraisal. Further weight is given to this view when exploring the relationship between roh-z values obtained for scale scores and the average of the roh-values for individual items

constituting the summary score. The idea was first formulated by O'Muircheartaigh (1977) and a methodological appraisal developed in chapter 7 following McKennell (1977) to examine whether or not summary scale scores could be 'interviewer effect free' by eliminating certain items with particularly high values of r_{oh} . The findings suggest that around half of the twelve summary scores based on the FLP could be redefined to be almost 'free' of any effect despite the fact that individual items would still be sensitive to interviewer variability. Whether or not the resulting subsets of items retain substantive coherence is debatable. Given changes in item sensitivity over time it would still seem wise to carry out such evaluations each time the schedule is administered in a survey. Multivariate appraisal of the 'dimensionality' of any interviewer effect replicated strategies illustrated in O'Muircheartaigh (1976) and O'Muircheartaigh and Wiggins (1981). For the two scales considered, namely 'sleep and rest' and 'recreation', there is strong evidence that whilst similar, interpretable subsets of items may be identified as being sensitive to any 'effect' over time these 'effects' may be generated by different subgroups of interviewers. There are enormous implications for further research, especially that which focuses on interviewing style, (e.g. Dijkstra, 1981). Attention should be placed on why certain interviewers appear 'stable' over time and why others are not. The impact of mispecifying 'constraints' in estimating interviewer effects based on a one way classification is also illustrated. Although, the results do not show any dramatic discrepancies the potential for misinterpretation remains. Using the structure of year one interviewer effects for the 'sleep and rest' scale to predict the nature of the following year's impact fails to be convincing, largely due to the observed volatility in the nature of the effects for individual items.

(b) Non response bias:

in chapter 13 estimates of non response bias at different stages in the calling process are combined with estimates of sampling variance and cost to produce an 'efficiency' criteria to examine whether or not results of similar accuracy could be achieved more cheaply. In terms of 'bias' alone, in the Southern domain of the OMS study most of the bias did arise during the initial call, but not so in the Northern domain. For both domains substantial reductions in bias were witnessed by the third stage of calling. Appointment calls looked especially attractive; the success rate

of appointment calls to non appointment calls was typically between 2.4 and 4.9. However, the attraction was much less appealing in terms of relative bias. These conclusions hold broadly for all three methods of bias estimation. The conclusion would appear to be; experiment with three versus four calls strategies, encourage use of appointments, but be wary of adopting call back norms like 'three calls then only follow up outstanding appointment calls'. For the global evaluation costs and sampling variance were combined with each method of estimation for each domain across six separate item evaluations. For both domains under cost model I (considered to be the least realistic with a constant travel cost assumption at each stage of calling) there was support for the 'status quo' strategy of making all four calls. Under the other cost models (II and III) the results diverge. In the North, wherever there was evidence to question the 'status quo' it was towards the three call norm, with some support for a two call norm for particular items. In the South, about half of the items considered, provided continued support for the 'status quo' under both cost models. One of the main attractions about the empirical evidence presented in part two is that it really does begin to question the wisdom of fixing callback norms 'a priori'. It also demonstrates an evaluative scheme whereby components of nonsampling and sampling error can be combined with cost as an index of 'quality and precision'. The main weaknesses of the strategy are its' dependence on modelling for costs and bias and the univariate character of the appraisal. Hopefully, the development of further methodological refinements in the future to remedy these deficiencies will motivate empirical studies to provide sound evidence on the wisdom of certain callback plans.

iii) Modelling context:

Variance component modelling (VARCL), reviewed in chapter 3 and illustrated in chapter 8, demonstrates a valuable new approach to the analysis of the impact of interviewer variability on the results of sample surveys. A framework is established within which traditional analysis of interviewer variance is embedded. Modelling relationships between survey variables is conducted in the presence of interviewer effects. The ANS study illustrated the treatment of a binary variable, and is a direct extension of the work by O'Muircheartaigh and Wiggins (1981). Analysis of the PHS study is based on a quantitative response measure (the overall FLP score) modelled in terms of several

survey variables for two separate waves. The analysis is also conducted for the combined waves, where the response for each wave is nested within each respondent; interviewers are included as factors in the explanatory part of the model. In all of the illustrations, despite strong evidence of interviewer variability (PHS wave two is the exception here) the substantive relationships only tend to be 'masked' by the presence of interviewers. What becomes exciting is the evidence about the direct inclusion of interviewer characteristics in the modelling. The findings are only limited by the scope of the measured characteristics and the number of interviewers included in the experiments. Nevertheless, in the ANS study about 50 percent of the variability introduced into the responses by the interviewers themselves seems to be attributable to variations in interviewer experience and response rates. In the first wave of the PHS study, the influence of the average number of calls and an interviewer's own attitude to the disabled seems to account for around 50 percent of the observed interviewer variability. In the second wave, only the average number of calls looked to be influential. For the combined analyses the different findings for each separate wave were confirmed by the presence of strong 'interviewer-year' interactions in the modelling. More importantly, the very direction of interviewer effects for individual interviewers seemed inconsistent between the waves. This confirmed the earlier findings based on univariate and multivariate appraisals. Only two interviewers, 4 and 5, appeared 'stable' in terms of the direction of their influence. Generally, different interviewers tend to behave differently for different items across the years. Again, confirming the view that 'once and for all' schedule evaluations in terms of interviewer variability could be misleading. Interestingly, some of the interviewer characteristics that looked to be influential often reflect fieldwork practice or outcomes, e.g. average number of calls per workload and interviewer response rate. Obviously with small numbers it is difficult to make any sweeping generalisation, but there is nothing inherent in the methodology to inhibit the inclusion of more subtle outcomes, like refusal rates as distinct from non-contact rates, or whether or not the response was the result of an appointment or nonappointment call. The illustrations in chapter 8 are useful because they bind the concerns of the methodologist with the data analyst by incorporating the presence of interviewers in the production of responses with the interpretation of the relationships between those responses in any substantive

analysis. Furthermore, there is a grand legacy of literature (e.g. Cannell, 1970) to guide the appropriate inclusion of interviewer characteristics in any modelling.

14.3:Future work

The preceding discussion has identified several features of the methodological investigation of nonsampling error that require more attention. Formally, it is possible to present these suggestions under four areas, response error, nonresponse error, costs and technological advances.

i) Response error:

it is important to conduct response error research jointly with investigators concerning nonresponse. Joint investigations are appropriate because securing a respondent's cooperation is the primary task assigned to the interviewer. The initial introduction (Morton-Williams and Young, 1986) is crucial and, if successful will set the 'tone' of any ensuing social interaction. Changing interviewer behaviours to increase respondent cooperation may also threaten the quality of responses obtained. It is clear from the results of replicating studies of interviewer variability for the same set of interviewers in the PHS study that issues around the nature of response error will not be answered by a single piece of research. Results confirm Kish's early work (1962) that no category of item can be thought 'a priori' to be 'interviewer effect free'. Different items may be sensitive to an effect for different subsets of interviewers across particular time intervals. In terms of quantitative assessment more work needs to be done to develop measures of interviewer variability for categorical items. Insights into the sources of interviewer variability will only be gained by closer inspection of interviewers working practices. This could involve videotaping, feedback sessions (see Miller and Cannell, 1982 and Cannell et al., 1981) and work on voice intonation (Barath and Cannell, 1976 and Oksenberg et al., 1986). This all suggests that cooperation between disciplines with mutual interests in methodology is needed (e.g. that between survey statisticians and cognitive psychologists). Experiments on question wording (see Kalton et al., 1978) also offer valuable information on sources of response error; these could be combined with work on

interviewer style (Dijkstra, 1987), effectiveness (Steinkamp, 1966) and technique (Cannell et al., 1979), barriers to recall (Moss and Goldstein, 1979), context effect (Brenner et al., 1978) and saliency (see Committee of National Statistics Research Council report, 1985). However, to believe that such courses of action will by themselves lead to 'standardised' delivery styles that will ultimately eliminate interviewer effect is naive. Their purpose will be to better inform the selection of interviewer characteristics to include in modelling sources of interviewer variability. There is enough empirical evidence contained in part one to suggest that no matter how much careful training, briefing and supervision interviewers receive errors will continue to arise in interviewing because both the researcher and the actual questions remain' potentially misinterpreted and misinterpreting' (Cicourel, 1964). Minimum design modifications for at least part of the sample appeal as pragmatic checks on the quality and precision of survey data. It is pragmatic to do so simply because interviewers are present during the data production process, so why shouldn't they be present during the estimation and analytical stages of data processing and interpretation? By the same token, experiments to estimate the influence of other sources of response error (e.g. coder variability, see Collins and Kalton, 1981) could also be included in any appraisal.

Replicated studies like the PHS are clearly an important advance over and above 'one shot' studies. However, it is not enough simply to replicate in the same mode of data collection. Comparison of errors in different modes of data collection would also be useful.

ii) Non response error:

in the absence of external validating information more emphasis will need to be placed on data analytic techniques both to assess and improve the quality of survey estimates. As many of the techniques reviewed and demonstrated in part two depend on there being a relationship between the value of a survey variable and the amount of effort required to obtain a response then this assumption needs closer empirical inspection. Such evidence can only be provided by routinely collecting and processing call back information, e.g. time of call etc. so that investigators can select 'imputation' models that properly match the context of data collection. More work like that of Thomsen and Siring

(1983) and Lievesley (1986) for continuous surveys will also help investigators re-appraise the assumptions underlying estimation procedures.

iii) Costs:

the following observation from Kish (1965) still reverberates around current survey practice....' ordinarily the sampler has no precise data on cost factors, and must base his decisions on estimates or guesses a good cost model helps to ask the right questions and to make good guesses....' The reporting of survey costs is all too rare in the survey literature. Results for three costs models are presented in chapter 13. It is hoped that they have reasonable intuitive appeal. A broader discussion as to what actually constitutes a 'good cost model' is essential if routine evaluations like the one in part two can be carried out. Combining 'good' cost information with mean square error would obviously lead to a better understanding of the survey process that might result in better informed budgets and reviews of call back strategies. An outstanding issue is the need to attempt multi-item appraisals. It is difficult to understand why Kish's global optima/proximal framework has not been applied to this goal (Kish, 1976, 1986).

iv) Technological advance:

some of the most exciting research potential for realising global evaluations of the sort demonstrated in this thesis may be created by recent technological developments. (for an interesting review of the implications of computer assisted techniques for survey organization, see survey methods Newsletter, Summer, 1989). 'Computer assisted personal interviewing' (CAPI) and 'computer assisted telephone interviewing (CATI) greatly reduce the coding and data processing work needed to produce a machine readable data base. As well as eliminating certain sources of nonsampling error these innovations could release more resources for methodological work on telephone interviewing or comparative research on mixed modes of data collection (e.g. see Dillman, 1978, Jordan et al., 1980, Tucker, 1983, Sykes and Hoinville, 1985 and the Market Research Society, 1986). Also given the concern about the absence of cost data in most evaluations it is encouraging to see evidence of research on cost models for CATI surveys (Bryant and Weidman 1987).

The progressive dilution in the proportion of face to face modes of data collection, especially in the USA, make methodological research on CATI systems highly relevant. With such systems interviewers assignments can be routinely randomised so that interviewer variance estimates can be provided for any study. By locating monitors at other terminals independent recordings of a respondent's answers could be made, together with video taping to provide immediate feedback on the quality of an interview. Groves and Kahn (1979) suggest that if particular questions are very important to the research effort then a monitoring system could be focussed on those questions intensively to allow changes in the form of wording during the period of study. Given the technological capabilities available detailed information about all management could be collected, e.g. precise control over the number of call attempts could be easily implemented as well as ideas like assigning low incidence calls a high priority in call back allocations. Routine interviewer profiles containing productivity ratings and overall item sensitivities could be made available on a daily basis.

However patterns of modes of data collection change it is hoped that the quality of survey research will benefit from the type of methodological appraisal exemplified in this thesis. Perhaps technological advance will mean that there will be fewer obstacles to experimentation in survey design so as to help unite features of data production with concerns of data analysis. Statisticians will then be able to take proper account of the conditions which shape their data. Design modification will become design 'accommodation'. After all, if '... chance, randomness and error constitute the very core of statistics, we statisticians must include chance effects in our patterns, plans, designs and inferences' (Kish, 1987)!!

BIBLIOGRAPHY

Aitken, M., Anderson, D., and Hinde, J. (1981). Statistical modelling on teaching styles. *J.R. Statist. Soc. A*, 144, Part A, pp. 419-461.

Aitkin, M., and Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, 149 (1), 1-43.

Altham, P.M.E. (1976). Discrete analysis for individuals grouped into families, *Biometrika* 63, 2, 263-9.

Anderson, D., and Aitken, M. (1985), Variance component models with binary response: interviewer variability, *JRSS(B)*, 47, No. 2, 203-10.

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research*, McGraw Hill, 1st ed.

Baker, R.J and Nedler, J.A. (1978). *The GLIM system Manual Release 3. Generalised interactive modelling*. London: Royal Statistical Society.

Barath, A and Cannell, C.F. (1976). Effect of interviewer's voice intonation. *Public Opinion Quarterly*, 40,3, pp. 370-373.

Bartholomew, D.J. (1961). A method for allowing for not at homes in sample surveys, *Applied Statistics*, 10, 52-9.

Bartlett, F.C. (1932). *Remembering*. London: Cambridge University Press.

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests, *Proc. Roy. Soc. London, Series A*, Vol. 160, 268-282.

Bead, T.W., and Stimson, R.J., eds. (1985). Survey Interviewing :
George Allen & Unwin.

Bell, J. (1985). Generalizability theory: the software problem, *Journal of Educational Statistics*, Spring 1985, Vol. 10, No. 1, 19-29.

Birnbaum, Z.W., and Sirken, M.G. (1950). Bias due to non-availability in sampling surveys, *J. Am. Stat. Assoc.*, **45**, 98-111.

Booker, H.S., and David, S.T. (1952). Differences in results obtained by experienced and inexperienced interviewers, *JRSS(A)*, **115**, 232-257.

Boyd, H.W., and Westfall, R. (1965). Interviewer bias revisited, *Journal of Marketing Research*, **2**, 58-63.

Brenner, M., ed., (1978) . *The Social Contexts of Method*: edited by M. Brenner, Peter Marsh and Marilyn Brenner, Croomhelm Ltd.

Brenner, M. (1980). The social psychology of the research interview symposium paper. University of Surrey (unpublished).

Broemling, L.D. (1969). Confidence regions for variance ratios of random models, *J. Am. Stat. Assoc.*, **64**, 660-664.

Bryant, K.J., and Weidman, L. (1987). Developing cost models for CATI Surveys. *Proceedings of the survey research methods section, of the Amer. Statistical Assocn.*

Bush, N., and R.L Anderson (1963). A comparison of three different procedures for estimating variance components. *Technometrics* **5**, 421-440.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in Social settings. *Psychol. Bull.*, **54**, 297-312.

Cannell, C.F. (1977). A summary of studies of interviewing methodology. Vital and Health Statistics. Series 2 - No. 69. Data evaluation and methods research. U.S. Dept. of Health, Education and Welfare. U.S. Govt. printing office Washington D.C. DHEW pub. No. (HRA) 77-1343.

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., and Fowler, F.J. (1965). Comparison of hospitalization reporting in three survey procedures. *Vital and Health Statistics, Series 2*, No. 8, U.S. Dept. of Health, Education and Welfare, Public Health Service.

Cannell, C. and Kahn, R.F. (1968). "Interviewing" in the *Handbook of Social Psychology*, Vol. 2, 2nd Ed., eds. G. Lindzey and E. Aronson, Addison Wesley, Reading, Mass.

Cannell, C.F., Marquis, K.H., and Laurent, A. (1970). A summary of research studies in interviewing methodology, Ann Arbor: Survey Research Center.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981). Research on interviewing techniques. In: Leinhardt, S. (ed), *Sociological Methodology 1981*. Jossey-Bass, San Francisco.

Chapman, D.W. (1974). An investigation of non-response imputation procedures for the health and nutrition examination survey: prepared for Division of Health Examination Statistics, Nat. Center for Health Statistics, HEW by Westat, Inc.

Chapman, D.W. (1976). A survey of non-response imputation procedures', proceedings of the social statistics section, American statistical Association, part II, 491-4.

Charlton, J. (1981). Sociomedical factors and level of disability in The Longitudinal disability interview survey, Phase 1 Report, January 1981, St. Thomas' Hospital Medical School.

Charlton, J., Patrick, D.L., H. Peach (1983). Uses of multivariable measures of disability in health surveys, *Journal of Epidemiology and Community Health*, 1983, 37, 296-304.

Choi, I.C., and Comstock, G.W. (1975). Interviewer effect on responses to a questionnaire relating to mood, *Amer. J. Epidemiology*, 101, 84-92.

Cicourel, A.V. (1964). *Method and measurement in sociology*. The Free Press, New York.

Clark, A.L., and P. Wallin (1964). The accuracy of husbands' and wives' reports of the frequency of marital coitus. *Population studies*, 18(2), 165-173.

Cochran, W.G. (1963). *Sampling Techniques*, 2nd ed., New York: John Wiley and Sons.

Cochrane, A., Chapman, P.J., Oldham, P.D., (1951). Observer errors in taking medical histories. *The Lancet*, 1007-9.

Collins, M. (1978). Interviewer variability: The North Yorkshire experiment, *Journal of Mkt. Res. Soc.*, Vol. 20, pp 59-72.

Collins, M. (1979). Interviewer variability: Milton Keynes household survey. *Methodological Working Paper*. No 16. London. SCPR.

Collins, M., (1981). Interviewer variability: a review of the problem, (1981). *J. of Mkt. Res. Soc.*, Vol. 22, No.2.

Collins, M., Butcher, B. (1986). Interviewer and clustering effects in an attitude survey, J. of Mkt. Res. Soc., Vol. 25, No.1.

Coombs, C.H. (1964). A theory of data. New York: Wiley.

Cobb, S. and Cannell, C.F. (1966). Some thoughts about interview data, Int. Epidemiological Bulletin, 13, 43-54.

Cox, B.G. (1980). The weighted sequential not deck procedure, Proceedings of survey research methods section American Statistical Association, Horston Arg. 11-14.

Cox, D.R. (1970). The Analysis of Binary Data, Chapman & Hall.

Cramer, H., Mathematical Methods for Statistics, (1946), Princeton Univ. Press.

Cronbach, L.J. (1946). Response sets and test validity. Educ. psychol measurement., 6, 475-494.

Cronbach, L.J. (1960). Essentials of psychological testing (2nd. ed.) New York: Harper and Row.

Curtis, S. (1983). Interviewer effects association with variations, Sec. 3.5, Vol. I, 48-55 of Intra-urban variations in health and health care: the comparative need for health care survey of Tower Hamlets and Redbridge.

Delanius, T. (1957). Sampling in Sweden, Stockholm, Almquist & Wicksell.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between resultant error of response and the bias of nonresponse, JASA.

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

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation of covariance components models. J. Amer. Statist. Ass., 76, 341-353.

Dempster, A.P. et al. (1977). Maximum likelihood for incomplete data via the EM algorithm. J.R. Statist. Soc. B 39, 1-38.

Dijkstra, W. (1987). Interviewing style and respondent behaviour: an experimental study of the survey-interview. Sociological methods and Research, Vol. 16, No 2, 309-334.

Dijkstra, W., and van der Zouwen, J., (1978). Role playing in the interview: towards a theory of artifacts in the survey-interview, in Sociocybernetics, Vo. 2, (eds. R.F. Geyer and J. van der Zouwen), pp. 59-83, Martinus Nijhoff, Leiden.

Dillman, D.A. (1978). Mail and Telephone Surveys: the total design method. Wiley, New York.

Dohrenwend, B.P. (1966). Social status and psychiatric disorder: an issue of substance and an issue of method', . Amer. Sociological Review, 31, 14-34.

Drew, J.H. Fuller, W.A. (1980). Modelling nonresponse in surveys with callbacks, Proceeding of Sampling Section of Am. Stat. Assocn.

Durbin, J. (1954).Nonresponse and callbacks in surveys, BISI, 3412, 72-86.

Durbin & Stuart, (1951). Differences in response rates of experienced and inexperienced interviewers, (1951), JRSS (A), 163-205.

Durbin, J., and Stuart, A. (1954). Call-backs and clustering in sample surveys. JRSS (A), 117, 387-428.

Eiden, Van der, (1989). Personal Communication.

Ericson, W.A., (1967). Optimal sample design with nonresponse, JASA, 62, 63-78

Eisenhart, C. (1947). The assumptions underlying the analysis of variance, Biometrics.

Everitt, B.S., (1977). The Analysis of Contingency Tables, . Chapman Hall.

Everitt, B.S., (1984). An introduction to Latent variable models. Chapman and Hall, London, New York.

Fairbairn, A.S., Wood, C.H., Fletcher, C.M., 'Variability in answers to a questionnaire on respiratory symptoms', (1959), Brit. J. of Preventative Social Medicine, 17.

Federer, W.T. (1955). Experimental Design, Macmillan, New York.

Feldman, J.J, Hyman, H and Hart, C.W. (1951). A field study of interviewer effects on the quality of survey data. Public Opinion Quarterly, 15, pp 734-761.

Fellegi, I.P. (1964). Response variance and its estimation, J. of American Statistical Association, 59, 1016-1041.

Flowerman, S.H., et al, (1950). Unpublished Report, American Jewish Committee.

Fowler, F.J., Jr. (1966). Education, interaction, and interview performance. Doctoral dissertation. University of Michigan.

Frankel, M.R., (1977), Inference from survey samples. Ann Arbor: Institute of Social Research, University of Michigan.

Frankel, L.R., Dutka, S. (1983). Survey design in anticipation of nonresponse and imputation, in Chapter 3, incomplete data in sample surveys, Vol. 3, Session II, Academic Press.

Freeman, M.F. and Tukey, J.W., 'Transformations related to the angular and square root transformations', (1950), Annals. Math. Stat., Vol. 21, 607-611.

Freeman, J. and Butler, E.W. (1976). Some sources of interviewer variance in surveys, Public Opinion Quarterly, Vol. 40, 79-91.

Gabriel, K.R. (1968). Simultaneous test procedures in multivariate analysis of variance, Biometrika, 55, 3, 489-503.

Gales, K.E., and Kendall, M.G. (1957). An inquiry concerning interviewer variability, JRSS (A), 69, 496-501.

Ganguli, M. (1941). A note on rested sampling, Sankhya, 5, 449-452.

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.

Goldberg, D.P., The Detection of Psychiatric Illness by Questionnaire, (1972), Oxford Press: London.

Goldstein, H., (1987). Multi level models in educational and social research: Griffin, London.

Goldstein, H., (1986). Multi level mixed linear model analysis using interactive generalised least squares. 73, pp 43-56.

Goldstein, H., (1988). Correspondence.

Goodman, L. (1973). Casual analysis of data from panel studies and other kinds of surveys, American Journal of Sociology, 82 (6),1289-1317.

Gorden, R.L. (1954). An interaction analysis of the depth interview. Doctoral dissertation, University of Chicago.

Gove, W.R., and Geerken, M.R. (1976). Response bias in surveys of mental health: and epirical investigation,American Journal of Sociology, 82 (6),1289-1317.

Gray, P.G.(1956). Examples of interviewer variability taken from two sample surveys, Applied Stats., 5, 73-85.

Gray, P., and Corlett, T. (1950). Sampling for the social survey, J. R. Stat. Soc. A, 113, 150-206.

Graybill, F.A. (1961). An Introduction to Linear Statistical Models: Vol. I, New York: McGraw Hill.

Groves, R.M., Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interviews, Academic Press.

Guest, L.L., (1947). A study of interviewer competence. *Int. J. opinion Attitude Research*, 1 (4), 17-30.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys, *JASA*, 41, 517-529.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961). Measurement error in censuses and surveys, *Bull. Int. Stats. Inst.*, 38, 359-374.

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

Hanson, R.H., and Marks, E.S., (1958). Influence of the interviewer on the accuracy of survey results', *JASA*, 53, 635-655.

Hartley, H.O., (1946). Contribution to the discussion of a paper by F. Yates, *J.R.S.S.*, 109, 34.

Hartley, H.O. (1967). Expectation, variances and covariances of ANOVA mean squares by 'synthesis'. *Biometrics*, 23, 105-114 and corrigenda, 853.

Hartley, H.O., and Rao J.N.K., (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

Harville, D.A. (1969). Quadratic unbiased estimation of variance components for the one way classification', *Biometrika*, 56, 313-326.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *JASA*, 72, 320-324.

Hatchett, S., and Schuman. J. (1975). White respondents and race of interviewer effects. *Public Opinion Quarterly*, 39, 523-528.

Hemmerle, W.J., and Hartley, H.O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the E transformation. *Technometrics*, 15, 819-831.

Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics*, 9, 226-252.

Henderson, C.R., (1959, 1969). Design and analysis of animal husbandry experiments: in Chapt. 1 of *Techniques and procedures in animal science research*, 1st ed. 1959, 2nd ed. 1969, American Society of Animal Science.

Henderson, C.R., Searle, S.R. and Schafer, L.R. (1974). The invariance and calculation of method 2 for estimating variance components. *Biometrics* 30, 583-588.

Herbach, L. H. (1959). Properties of type II analysis of variance tests, *Ann. Math. Stat.*, 30, 939-959.

Hildum, D., and Brown, R.W. (1956). Verbal reinforcement and interviewer bias. *J. of Abnormal Social Psychology*, 53, 108-111.

Hochstim, J.R. (1967). A critical comparison of three strategies of collecting data from households', *J. Am. Stat. Assoc.*, 62, 976-89.

Hochstim, J.R., Athanasopoulos, D.A. (1970). Personal follow-up in a merit survey: its contribution and cost, *Public Opinion Quarterly*, 34, 68-81.

Holland, W., Ashford, J.R., Colley, J.R.T., et al. (1966). A comparison of two symptom questionnaires, *Brit. J. of Preventive Social Medicine*, 20, 76-96.

Hyman, H.H. (1954). *Interviewing in Social Research* Chicago: University of Chicago Press.

Janis, I.L. (1958). Psychological Stress, , New York: Wiley.

Johnson, F.J. (1978). The price and relevance of accuracy of market research survey data, J. Mkt. Res. Soc., Vol. 25, No. 2.

Jordan, L.A., Marcus, A.C., and Reeder, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. Public Opinion Quarterly, 44 (2) pp. 210-222.

Joreskog, K.G., and Sorbom, D, (1986).

LISREL VI; Analysis of linear structural relationships by the method of maximum likelihood; users guide, 1986, University of Uppsala, Department of Statistics PO Box 513, S-751 20 Uppsala, Sweden

Jowell, R., and Airey, G. (1986). British Social Attitudes Survey: The 1986 Report, , Gower Press.

Kahn, R.L., and Cannell, C.F. (1957). The Dynamics of Interviewing: Theory, Technique and Cases, New York: Wiley.

Kalton, G., and Stowell, R. (1979). A study of coder variability. Applied Statistics, 28,3, pp. 276-289.

Kalton, G., Roberts, J., Holt, D. (1979). The effects of offering a middle response option with opinion questions , The Statistician, 29, Vol. , No. 1. 65-78.

Kempthorne, O. (1975). Fixed and mixed models in analysis of variance. Biometrics 31, 473-486

Kemsley, W.F.F. (1979). Interviewer vairability in expenditure surveys', (1979), JRSS (A), 128 , 118-139.

Kendall, M.G., and Stuart, A. (1961). The Advanced Theory of Statistics, Vol. 33, Chapter 33, Griffin: London.

Kim, J.I., and Fluek, J.A. (1976). A review of randomized response models and some new results. Proceedings of the Annual Meeting of the American Statistical Association, Social Statistics section pp. 477-482.

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

Kish, L. (1962). Studies of interviewer variance for attitudinal variables, , JASA, 57, 92-115.

Kish, L. (1976). Optima and proxima in linear sample designs, JRSS Soc. A, 139, Part I.

Kish, L. (1986). Personal Communication.

Kish, L. (1987). Statistical design for research: Wiley.

Kish, L., and Frankel, M.R. (1970). M.R., Balanced repeated replications for standard errors, JASA, Vol. 65, No. 331.

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

Kish, L. and Hess, I. (1959). A 'replacement' procedure for reducing the bias of nonresponse. The American Statistician, 13,4, 17-19.

Kish, L. and Lansing, J. (1954). Response errors in estimating the value of homes, JASA, 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, , J. Consulting and Clinical Psychology, 43(4), 448-452.

Krasner, L. (1958). Studies of the conditioning of verbal behaviour. Psychol. Bull., 55, 148-170.

Krewski, D. and Rao, J.N.K. (1981). Influence from stratified samples: properties of the linearization, Jack Knife and Balanced Repeated Replications Methods. Annals of Statistics 9, 1010-1019.

Koch, G.G. (1968). Some further remarks on "a general approach to the estimation of variance components", Technometrics, 10, 551-58.

Koch, G.G. (1967). A general approach to the estimation of variance components, Technometrics 9, 93-118.

Kokan, A.R., and Kahn, S. (1967). Optimum allocation in multivariable surveys: an analytical solution, J. R. Stat. Soc., B, 115-125.

Lamale, Helen. H. (1959). Methodology of the survey of consumer expenditures in 1950. Philadelphia: Univ. of Pennsylvania.

LaMotte, L.R. (1972). Notes on the covariance matrix of a random nested ANOVA model. Ann. Math. Statist. 43, 659-662.

Lansing, J.B., G.P Ginsberg and Kaisa Braaten. (1961). An investigation of response error. Bureau of Economic and Business Research, Univ. of Illinois.

Lawley, D.N and Maxwell, A.E. (1971). Factor Analysis as statistical method. Butterworths.

Leone, F.C., Nelson, L.S., Johnson, N.L. (1968). Eisenstat, S., Sampling distributions of variance components II empirical studies of unbalanced rested designs. *Technometrics* 10, 719-738.

Lievesley, D. (1986). Unit nonresponse in interview surveys, Unpublished.

Little, R.J. (1987). Models for non response in sample surveys, *JASA*, 77, pp 237-50.

Locker, D., Wiggins, R.D., Sittampalam, Y., and Patrick, D. (1981). Estimating the prevalence of disability in the community: the influence of sample design and response bias, *J. of Epidemiology and Community Health*, 35, No. 3, 208-212.

Longford, N.T. (1986). Variance components as a method for routine regression analysis of survey data. In *proceedings of Compstat 86*, Rome. Eds. De Antoni F., Lauro N. and Rizzi A. Physica-Verlag, Heidelberg, Vienna.

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. (1988a). VARCL Manual, Educational Testing Service, Princeton, N.J., U.S.A.

Longford, N.T. (1988b). Normal variance component modelling and exponential family extensions. Proceeding of the section on statistical computing of the American Statistical Association, Alexandria, VA, 137-142.

Longford, N.T. (1988c). Correspondence.

McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys, National Centre for Health Statistics, Series No. 2, No. 14, Washington U.S. Dept. of Health, 1966.

Macfarlane-Smith, J. (1972). Interviewing in Market and Social Research, Routledge, Kegan & Paul.

McGeoch, J.A. (1932). Forgetting and the law of disuse. *Psychol, Rev.* 39: 352-370

McGowan, I., (1986). Fitting S shaped curves to mail response data, *The Statistician*, Vo. 35, No. 1.

McKennell, A. (1973). Psychosocial factors in aircraft noise annoyance, In *Proceedings of the International Congress on Noise as a Public Health Problem*, (Dubrovnik), 627-644, U.S. Environmental Protection Agency Publication 550/973 008: Washington.

McKennell, A.C. (1974). *Survey Attitude Structures*, pp 45-55. Amsterdam: Elsevier.

McKennell, A.,(1977). *Attitude Scaling*, Chapter 8, Vol. 2, *Model Fitting: The Analysis of Survey Data*, Wiley.

McKenzie, J.R. (1977). An investigation into interviewer effects in market research. *Journal of Marketing Research*, 14, pp 330-336.

Madow, W.G., Nisselson, H., and Olkin, I. (1983). *Incomplete Data in Sample Surveys*, Volume 1, Report and Case Studies: New York: Academic Press.

Madow, W.G., and Olkin, I. (1983). *Incomplete data Sample Surveys*, Volume 3, proceedings of the symposium. New York: Academic Press.

Madow, W.G., and Olkin, I., and Rubin, D.B. (1983). Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies, New York: Academic Press.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J of the Royal Statistical Society*, 109, pp 325-370.

Mandler, G., and Kaplan, W.K. (1956). Subjective evaluation and reinforcing effect of a verbal stimulus. *Science*, 124, 582-583.

Mantel, N., 'On rapid estimation of standard errors for the means of small samples', (1951), *Amer. Statistician*, 5, 26-27.

Market Research Society Development Fund. (1986). Comparing telephone and face to face interviews: a report on a methodological research survey.

Manson, W.M., Wong, G.Y. and Entwistle, B. (1984). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology 1983-1984* (pp. 72-103). San Francisco: Jossey-Bass.

Maxwell, A.E. (1977). *Multivariate analysis in behavioural research*: Chapman and Hall.

Maximum Likelihood Program. (1985). Release 3.08, issued by 'The Numerical Algorithms group Ltd, NAG Central Office, Oxford, United Kingdom.

Miller, P.V., and Cannell, C.F. (1982). A study of experimental techniques for telephone interviewing. *Public Opinion Quarterly*, 46, 2, pp. 250-269.

Morton Williams, J.(1979). The use of verbal interaction coding for evaluating a questionnaire, *Quality and Quantity*, 13, 59-75.

Morton Williams, J., and Young, P. (1986). Interviewer strategies on the doorstep, *Mkt. Research Society Proceedings of 29th Anrat Conference*.

Murthy, M.N. (1983). A framework for study incomplete data with a reference to the experience in some countries of Asia and the Pacific. Chapter 1, *Incomplete Data in Sample Surveys*, Vol.3. Academic Press.

NAG Forran Library Manual - Mark 12. (1987). Numerical Algorithms Group Ltd, NAG Central Office, Oxford.

Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews, *JASA*, 59, 18-55.

Oksenberg. L, Coleman. L. and Cannell, C.F. (1986). Interviewer's voices and refusal rates in telephone surveys. *Public Opinion Quarterly*, 50, 1, pp. 97-111.

O'Muircheartaigh, C.A. (1976). Response errors in an attitudinal survey, *Quality and Quantity*, 10, 97-115.

O'Muircheartaigh, C.A. (1976). The structure of interviewer effects, Unpublished.

O'Muircheartaigh, C.A., and Payne, C. (1977) Response errors, Chapter 7, *Model Fitting: the analysis of survey data*, Wiley.

O'Muircheartaigh, C.A., and Wiggins, R.D. (1977). Sample design and evaluation for an occupational mobility study, *Econ. and Social Review*, Vol. 8, No. 2, 101-115.

O'Muircheartaigh, C.A., and Wiggins, R.D. (1981). The impact of interviewer variability in an epidemiological survey, *Psychological Medicine*, 11, 817-824.

Osborne, V. (1989). The field information system. *Survey methodology Bulletin*, Social Survey Division, office of Population Censuses and Surveys.

Pannekoek, J. (1988). Interviewer variance in a telephone survey. *J. of Official Statistics* Vol.4. No, 4, pp. 375-384.

Pannekoek, J. (1989). Personal communication.

Parry, H.J., and Crossley, M.J. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 14, 61-80.

Patrick, D.L. (1981). Health and care of physically disabled in Lambeth. *The longitudinal disability survey, Phase 1 Report*, Dept. of Community Medicine, St. Thomas' Hospital Medical School, SE1.

Patrick, D.L. (1981). Screening for disability in an inner city, *J. of Epidemiology and Community Health*, Vol. 35, No. 1, 65-70.

Patterson, H.D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554.

Payne, C. (1977). The Log Linear Model for contingency tables, chapter 4 in Vol.2. *The Analysis of Survey Data*, edited by C. Payne and C.A. O'Muircheartaigh, Wiley.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation, *Draper's Co. Memoirs, Biometric Series*, No. 1, London.

Phillips, D., and Clancy, K. (1972). Some effects of social desirability in survey studies, Amer. J. of Sociology, 77, 921-940.

Platek, R., Singh, M.P., Tremblay, V. (1977). Adjustments for nonresponse in surveys, Survey Methodology Journal, Statistics Canada, Vol. 3, No. 2, 1-24.

Platek, R., and Gray, G.B. (1979). Methodology and application of adjustments for nonresponse, Proceedings of 42nd Session of International Statistical Institute Manila, Dec. 1979.

Politz, A., and Simmons, W. (1949). An attempt to get the 'Not-at-homes' into the sample without call-backs. J.A.S.A., 44, 9.

Politz, A., and Simmons, W. (1950). A note on 'an attempt to get the not at homes' into the sample without call-backs. J.A.S.A, 45, 136-7.

Postman, L. (1961). The present status of interference theory, in C.N. Coefer, ed., Verbal Learning and verbal Behaviour. New York. McGraw-Hill. pp 152-179.

Powney, J. and Watts, M. (1987). Interviewing in educational research Routledge and Kegan Paul Ltd.

Pritzker, L., Ogus, J., Hansen, M.H. (1965). Computer editing methods - some applications and results, Bulletin of the International Statistical Institute, Proceedings of 35th Session, Belgrade, Sept. 1965, 417-441.

Proctor, C. (1977). Two direct approaches to survey nonresponse: estimating a proportion with callbacks and allocating effort to raise the response rate, Proc. Social. Statistics Section, ASA, 284-290.

Quay, H. (1959). The effect of verbal reinforcement on the recall of early memories. *J. of Abnormal Social Psychology*, 59, pp. 254-257.

Quenouille, M.H. (1959). *Rapid Statistical Calculations*, New York: Hafner.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, New York: Wiley.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*, New York: Wiley.

Raudenbush, S.W. (1988). Educational Applications of linear models: a review, *Journal of Educational statistics*, Vol. 13, No 2, pp. 85-116

Richardson, S.A., Barbara. S. Dohrenwend, and D. Klein. (1965). *Interviewing*. New York: Basic Books.

Richardson, S.A., Hastorf, A.H, and Dornbusch, S.M. (1964). Effects of physical disability on a child's description of himself. *Child Development*, 35, pp 893-907.

Riesman, D. (1958). Some observations on interviewing in the teacher apprehension study. In P.F. Lazarsfeld and W. Thielsen (Eds), *The Academic Mind*. Glencoe, IL: Free Press.

Robson, S.R., (1957), A Personal Communication to H.R. Searle.

Robinson, J. (1965). The distribution of a general quadratic form in normal variables, *Australian J. of Stat.*, 7, 110-114.

Rochon, J., Kalsbeek, W.D. (1983). Variance estimation from multi stage sample survey data: the jackknife repeated replicate approach, Univ. North Carolina, SAS U.S. Users Conference Proceedings.

Rogers, C.R. (1951). Client centred therapy. Boston: Houghton Mifflin.

Ross, G. (1983). A personal communication.

Rubin, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys, Proc. of 42nd. Session of Int. Stat. Institute, Manila, Dec. 1979.

Rudan, J.W. and Searle, S.R. (1971). Large sample variances of maximum likelihood estimators of variance components in the 3-way nested classification, random model, with unbalanced data. *Biometrics* 27, pp. 1087-1091.

Salzinger, S. (1956). Rate of affect response in schizophrenics as a function of three types of interview verbal behaviour. Paper read at Eastern Psychol. Assoc. meeting, Atlantic City.

Salzinger, K., and Pisoni, S. (1960). Reinforcement of verbal affect responses of normal subjects during an interview. *J. of Abnormal Social Psychology*, 60, pp127-130.

Sandstrom, R. (1977). The reaction of refusal nonresponse - an attempt to design a model and its application, ISI Paper, New Delhi.

Saslow, G., et al. (1957). Test-retest stability of interaction patterns during interviews conducted one week apart. *J. Abnormal Social Psychology*, 54, pp. 295-302.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components, (*Biometrics Bulletin*, 2, 110-114.

Scheffe, H. (1965). The Analysis of Variance, Wiley.

Searle, S.R. (1970). Large sample variances of maximum likelihood estimators of variance components. *Biometrics* 26, pp. 749-788.

Searle, S.R., Linear Models, (1971), New York: Wiley.

Searle, S.R. (1986). Personal Communication.

Searle, S.R. (1987). Linear models for unbalanced data. New York: Wiley.

Sellitz, C., M. Jahoda, M. Deutsch, and S.W. Cook. (1959). Research methods in social relations. New York: Holt.

Shapiro, S., and Eberhart, J.C. (1947). Interviewer differences in an intensive interview survey. *Int. J. of Opinion Attitude Research.*, 1 (2), 1-17.

Shapiro, G., and Schevren, F. (1979). A hot deck application of multiple imputations in the cps income supplement, Unpublished Working Paper of Social Security Administration, Washington, D.C.

Sharp, H., and Feldt, A. (1959). Some factors in a probability survey of a metropolitan community. *American Sociological Review*, 24, 650-661.

Simmons, W.R. (1954). A plan to account for 'not at homes" by combining weighting and call-backs. *Journal of Marketing*, 19, 42-53.

Slater, M., and Wiggins, R.D. (1978). Interviewer workload allocation: with particular reference to studies of interviewer variability, Unpublished.

Speed, F.M. (1966). A new approach to the analysis of linear models, PhD Thesis, Texas A & M University College Station, Texas.

Steeth, C.G. (1981). Trends in nonresponse rates in 1952-1979, *Public Opinion Quarterly*, Vol. 45.

Steel, R.G.D., and Torrie, J.H. (1980). Principles and processes of statistics: a biometrical approach. McGraw Hill, New York.

Steinkamp, S.W. (1966). Some characteristics of effective interviewers, Journal of Applied Psyc. Vol. 50, No. 6, 487-492.

"Student", (1942), 'The distribution of means of samples which are not drawn at random', Biometrika, 7, 210-215.

Sudman, S., (1961). Probability sampling with quotas, JASA, Sept. 1961.

Sudman, S., Bradburn, N.M. (1974). Response Effects in Surveys, Chicago: Aldine Publishing Co.

Sukhatme, P.V., and Sukhatme, B.V. (1970). Sampling Theory of Surveys with Applications 2nd ed., Iowa State University Press.

Summers, G.F., and Beck, E.M. (1973). Social status and personality factors in predicting interviewer performance,

Survey Methods Seminar, (1985). Telephone interviewing, Various Contributors, SCPR Survey Methods Newsletter.

Survey Methods Centre Newsletter. (1989). Computer Assisted Survey Systems. Summer 1989, SCPR, London.

Sykes, W., and Hoinville, G. (1985). Telephone interviewing on a survey of social attitudes: a comparison with face to face procedures. Social and Community Planning Research, London.

Taylor, C. (1977). Principal Components Analysis, . Chapter 4, in Exploring Data Structures: Survey data analysis, Vol. 1, (eds. O'Muricheartaigh, C.A., and Payne, C.).

Taylor, C. (1987). Personal Communication,

Thompson, R. (1975). A note on the W transformation. *Technometrics*, 17, 511-512.

Thompson, R. (1980). Maximum likelihood estimation of variance components. *Math. Oper. Statist., Ser. Statistics* 11, 545-561.

Thompson, W.A., Jr.(1961). Negative estimates of variance components: an introduction, *Bulletin, Int. Inst. of Statistics*, 34, 1-4.

Thompson, W.A., Jr. (1962). The problem of negative estimates of variance components, *Ann. Math. Stat.*, 33, 273-289.

Thomsen, I., and Siring, E. (1983). On the causes and effects of nonresponse: Norwegian experiences', Chapter 2, *Incomplete Data In Surveys*, Academic Press, Vol. 3, Session I.

Tiao, G.C., and Box, G.E.P. (1967). Bayesian analysis of a three component hierarchical design model, *Biometrika*, 54, 109-125.

Townsend, E.C., Unbiased estimators of variance components in simple unbalanced designs, PhD Thesis, Cornell Univ. Ithaca, New York.

Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly* vol. 47; pp. 84-95.

Urquhart, N.S., Weeks, D.L., Henderson, C.R., (1970). Estimation associated with linear models: a revisit', Paper BU-195 Biometrics Unit, Cornell University, Ithaca.

Van Westerhoven, E.M.C. 1978. Covering nonresponse: does it pay - a study of refusers and absentees, *J. of Mkt. Research Society*, Vol. 20, No. 4.

Wang, Y.Y. (1967). A Comparison of several variance component estimators, *Biometrika*, 54, 301-305.

Waksber, G.J., and Pearl, R. (1965). New methodological research on labour force measurements, *Proc. Am. Statistical Assoc. Social Stats. Section*, 1965, 227-237.

Weeks, Jones, Folsom, and Benrud. (1980). Optimal times to contact sample households, *Public Opinion Quarterly*, Vol. 44.

Weiss, C., Validity of welfare mothers' interview responses. (1968). *Public Opinion Quarterly*, 32, 662-633.

Weiss, D.J., Davis, R.V, England, G.W. and Lotquist, L.H. (1961). Validity of work histories obtained by interview, *Minn. Stud. Vocational Rehabilitation*, 12, No. 41.

Wiggins, R.D. (1980). Sample design for West London survey of aircraft noise, in *Aircraft Noise and Psychiatric Morbidity* (ed. A. Tamopolsky and J. Morton-Williams), 12-29, SCPR: London.

Wiggins, R.D. (1981). A profile of interviewers attitudes to disability in a study of physical handicap, Unpublished.

Wiggins, R.D. (1985). A replicated study of the impact of interviewer variability in a community survey of phys. hand. in an Inner London Borough, *Research Working Paper*, No. 24, PCL.

Wiggins, R.D. (1988). A retrospective evaluation of a call-back strategy in a survey of occupational mobility. *Research Working Paper*, No.33, The Polytechnic of Central London.

Wiggins, R.D., Longford, N., and O'Muircheartaigh, C.A. (1990). A variance components approach to interviewer effects. Joint SCPR LSE Centre for Survey Methods Working paper No.2.

Wilk, M.B., and Kempthorne, O., Fixed, mixed or random models, (1955). J. Am. Stat. Assoc., **50**, 1144-1167.

Wilk, M.B., and Kempthorne, O. (1955). Some aspects of the analysis of factorial designs in a completely randomized design, Ann. Math. Stat., **27**, 950-985.

Williams, J.A., Jnr., (1964). Interviewer-respondent interaction: a study of bias in the information interview, Sociometry, **27**, 338-352.

Williams, J.S. (1962). A confidence interval for variance components, Biometrika, **49**, 278-281.

Wolter, K.M. (1985). Introduction to variance estimation, Springer verlag, New York.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. J Amer. stat. Assoc. 29, 51-66.

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

Yuker, H.E., Block, J.E., and Young, J.E. (1970). The measurement of attitudes towards disabled persons, Rehabilitation Series 3, Insurance Company of North America, Ina Mend Institute at Human Resources Centre, New York.

APPENDIX

A univariate assessment of items comprising the functional limitations profile for ten categories¹

¹ Footnote: Information for "Sleep and Rest" and "Recreation" is contained in table 7.8 and 7.9

Table A1: Eating Domain

item	Roh		Sample estimate	
	Wave 1	Wave 2	Wave 1	Wave 2
82 I am eating much less	52	-04	.19	.11
83 I feed myself but only by eating specially prepared food or by using special utensils	24	-14	.05	.03
84 Soft food, bland, low salt, low fat, or low sugar diet	13	6	.18	.17
85 I am eating no food at all but I am taking liquids	29	18	.01	.01
86 I just pick or nibble at my food	63	18	.07	.05
87 I am drinking less fluids	6	13	.05	.03
88 I feed myself with help from someone else	-12	15	.01	.02
89 I do not feed myself at all, but must be fed	-30	28	.00	.01
90 I am eating no food at all, except by tubes or intravenous infusion	0	0	.00	.00

 $r_{roh12} = -.34$

Table A2: Body Care Domain

item	Roh Wave 1	Roh Wave 2	Sample estimate	
			Wave 1	Wave 2
91 I make difficult moves with help, for example, getting into or out of cars	9	35	.20	.20
92 I do not move into or out of bed or chairs by myself but I am moved by a person or mechanical aid	47	51	.06	.06
93 I stand only for short periods	33	10	.32	.27
94 I do not keep my balance	-16	-10	.19	.18
95 I move my hands or fingers with some limitation or difficulty	25	0	.09	.08
96 I stand up only with someone's help	4	57	.02	.01
97 I kneel, stoop or bend down only by holding onto something	4	-20	.28	.26
98 I am in a restricted position all the time	12	20	.05	.04
99 I am very clumsy	34	20	.10	.08
100 I get in & out of bed or chairs by grasping something for support or using a stick or walking frame	-18	-32	.16	.16
101 I stay lying down most of the time	2	-45	.04	.01
102 I change position frequently	37	28	.11	.10
103 I hold onto something to move myself around in bed	90	42	.06	.06
104 I do not bathe myself completely, for example, I require assistance with bathing	23	25	.05	.07
105 I do not bathe myself at all, but am bathed by someone else	-1	-20	.02	.03
106 I use a bedpan with assistance	0	57	.00	.01
107 I have trouble getting on my shoes, socks or stockings	-1	-14	.18	.14
108 I do not have control of my bladder	45	212	.07	.07
109 I do not fasten my clothing, for example I require assistance with buttons zips or shoelaces	9	28	.06	.04

Cont/....

item	Roh	Wave	Sample estimate	
	1	2	Wave 1	Wave 2
110 I spend most of the time partly undressed or in pyjamas	- 1	31	.03	.01
111 I do not have control of my bowels	-20	-03	.02	.03
112 I dress myself, but do so very slowly	- 7	35	.14	.15
113 I get dressed only with someone's help	1	67	.03	.02
	$r_{roh12} = -.42$			

Table A3: Ambulation Domain

item	Roh Wave 1	Wave 2	Sample estimate	
			Wave 1	Wave 2
114 I walk shorter distances or often stop for a rest	9	0	.37	.43
115 I do not walk up or down hills	45	41	.28	.30
116 I use stairs only with a physical aid, for example a hand rail, stick or crutches	4	31	.27	.25
117 I go up and down stairs only with assistance from someone else	11	-23	.03	.02
118 I get around in a wheelchair	-20	-9	.02	.02
119 I do not walk at all	76	0	.02	.01
120 I walk by myself with some difficulty, for example I limp, wobble, stumble, I have a stiff leg	-23	-42	.20	.19
121 I walk only with help from someone	-10	19	.01	.02
122 I go up and down stairs more slowly, for example one step at a time or I often have to stop	-10	-8	.35	.35
123 I do not use stairs at all	49	62	.03	.03
124 I get around only by using a walking frame, crutches, stick, walls or holding on to furniture	2	-32	.10	.11
125 I walk more slowly	34	-9	.39	.43

$$\text{Roh}_{12} = .48$$

Table A4: Mobility Domain

item	Roh Wave 1	Roh Wave 2	Sample estimate	
	Wave 1	Wave 2	1	2
126 I am getting around only at home	-5	43	.03	.04
127 I stay in one room	61	25	.02	.01
128 I am staying in bed more	40	1	.05	.02
129 I am staying in bed most of the time	-12	57	.01	.01
130 I am not now using public transport	-11	1	.10	.12
131 I stay at home most of the time	20	-7	.22	.17
132 I go out if there is a lavatory nearby	62	-17	.05	.04
133 I am not going into town	21	-21	.18	.06
134 I stay away from home for short periods only	8	-43	.18	.17
135 I do not get around in the dark or in unlit places without someone's help	3	-4	.13	.27

$$r_{\text{roh}_{12}} = -.26$$

Table A5: Work Domain

item	Roh Wave 1	Roh Wave 2	Sample estimate	
	Wave 1	Wave 2	Wave 1	Wave 2
271 I am doing part of my job at home	0	0	.00	.00
272 I am not getting as much done as usual	-15	28	.03	.01
273 I often act irritable with my work mates, for example, I snap or criticise them easily	-13	-66	.02	.00
274 I am working shorter hours	-26	0	.02	.01
275 I am doing only light work	-13	104	.05	.03
276 I work only for short periods of time or rest often	-21	78	.03	.03
277 I am working at my usual job but with some changes, for example, I use different tools or special aids or I swap jobs with other workers	- 6	- 2	.01	.01
278 I do not do my job as carefully and accurately as usual	-21	- 69	.02	.01
281 Not working or retirement due to health	- 4	-3, -4 ⁺	.29	.25

$r_{\text{roh}12} = -.01$

⁺ composite based on two variables retirement or not working due to health

Table A6: Household Management Domain

item	Roh	Wave	Sample estimate	
	Wave 1	Wave 2	Wave 1	Wave 2
305 I do housework or work around the house only for short periods of time or rest often	55	-44	.32	.31
306 I am doing less of the regular daily work at home than I would usually do	2	-13	.28	.29
307 I am not doing any of the regular daily work at home that I would usually do	51	-26	.07	.04
308 I am not doing any of the maintenance or repair work that I would usually do in my home or garden	35	-11	.25	.23
309 I am not doing any of the shopping that I would usually do	-18	-4	.16	.10
310 I am not doing any of the housework that I would usually do	-36	24	.11	.10
311 I have difficulty working with my hands, for example, turning taps, using kitchen gadgets, sawing or repairs	23	99	.11	.06
312 I am not doing any of the clothes washing I would usually do	-13	-29	.07	.08
313 I am not doing any heavy work around the house	-2	1	.37	.35
314 I have given up taking care of personal or household business affairs, for example, paying bills, banking household accounts	-2	-8	.07	.03

$$r_{roh_{12}} = -.20$$

Table A7: Communication Domain

item	Roh	Wave	Sample	estimate
	Wave	1	Wave	1
398 I have trouble writing or typing	- 4	14	.12	.10
399 I communicate by gestures, for example, I nod my head, point, or use sign language	1	25	.00	.01
400 My speech is understood only by a few people who know me well	-11	6	.01	.01
401 I often lose control of my voice when I talk, for example, my voice gets louder or softer, trembles, changes unexpectedly.	-11	32	.07	.03
402 I don't write except to sign my name	13	3	.05	.03
403 I carry on conversation only when very close to other people or looking directly at them	0	101	.02	.01
404 I have difficulty speaking, for example, I get stuck for words, I stutter, I stammer, I slur my words	6	-22	.05	.02
405 I am understood with difficulty	-16	-27	.02	.01
406 I do not speak clearly when I am under stress	16	22	.05	.03

$$r_{roh_{12}} = -.06$$

Table A8: Alertness Domain

item	Roh		Sample estimate	
	Wave 1	Wave 2	Wave 1	Wave 2
407 I am confused and start to do more than one thing at a time	-24	4	.09	.07
408 I have more minor accidents, for example, I drop things, I trip and fall, I bump into things	9	73	.13	.11
409 I react slowly to things that are said or done	2	0	.10	.07
410 I do not finish things I start	-32	- 9	.09	.04
411 I have difficulty reasoning and solving problems, for example, making plans, decisions, learning new things	12	15	.13	.06
412 I sometimes behave as if I were confused, for example, I do not know where I am, who is around or what day it is	6	-21	.08	.03
413 I forget a lot, for example, things that happen recently, where I put things, or keeping appts	- 4	78	.10	.08
414 I do not keep my attention on any activity for long	68	27	.09	.07
415 I make more mistakes than usual	31	13	.11	.07
416 I have difficulty doing things which involve concentration and thinking	28	62	.15	.09

$$r_{roh_{12}} = .25$$

Table A9: Emotion Domain

item	Roh Wave 1	Roh Wave 2	Sample estimate	
	Wave 1	Wave 2	Wave 1	Wave 2
421 I say how bad or useless I am, for example, I run myself down, I swear at myself for things that happen	16	16	.09	.08
422 I laugh or cry suddenly	121	27	.08	.11
423 I often moan and groan because of pain or discomfort	78	43	.09	.11
424 I have attempted suicide	34	19	.01	.01
425 I act nervous or restless	38	-15	.16	.20
426 I keep rubbing or holding areas of my body that hurt or are uncomfortable	66	88	.14	.17
427 I act irritably and impatiently, for example I run myself down, I swear at myself, I blame myself for things that happen	51	32	.17	.16
428 I talk hopelessly about the future	29	9	.12	.12
429 I get sudden frights	25	0	.08	.08

$$r_{\text{roh}_{12}} = .45$$

Table A10: Social Interaction Domain

item		Roh	Wave	Sample estimate	
		Wave 1	Wave 2	Wave 1	Wave 2
430	I go out to visit people less often	23	-17	.29	.25
431	I do not go out to visit people at all	8	-20	.10	.11
432	I show less interest in other people's problems, for example, I don't listen when they tell me about their problems, I don't offer to help	9	2	.09	.06
433	I am often irritable with those around me, for example, I snap at people or criticise easily	-24	-26	.13	.15
434	I show less affection	- 8	23	.05	.09
435	I take part in fewer social activities than I used to, for example, I go to fewer parties or social events	22	-10	.27	.26
436	I am cutting down the length of visits with friends	-29	-33	.16	.14
437	I avoid having visitors	14	25	.06	.05
438	My sexual activity is decreased	7	75	.20	.13
439	I often express concern over what might be happening to my health	83	22	.22	.24
440	I talk less with other people	5	-25	.11	.06
441	I make demands on other people, for example, I insist that they do things for me or tell them how to do things	-21	-24	.03	.03
442	I stay alone much of the time	- 2	6	.14	.16
443	I am disagreeable with my family, for example, I act spitefully or stubbornly	-22	-30	.03	.03
444	I frequently get angry with my family, for example, I hit them, scream or throw things at them	35	-29	.02	.01
445	I isolate myself as much as I can from the rest of my family	- 6	101	.01	.01
446	I pay less attention to the children	-15	-24	.01	.01
447	I refuse contact with my family, for example, I turn away from them	0	-50	.00	.00
448	I do not look after my children or family as well as I usually do	-19	11	.05	.03
449	I do not joke with members of my family as much as I usually do	22	1	.07	.03

$$r_{\text{roh}_{12}} = .16$$