The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision

Colm O'Muircheartaigh (London School of Economics, UK) and
Pamela Campanelli (Survey Methods Centre at SCPR, UK) [1]

Colm O'Muircheartaigh, LSE, Houghton St, London WC2A 2AE, UK

Key Words: Response variance, interviewer effect, interviewer variance, multilevel models

1. INTRODUCTION

The interviewer is seen as one of the principal sources of error in data collected from structured face-to-face interviews. Survey statisticians have expressed the effect in formal statistical models of two kinds. In the analysis of variance (ANOVA) framework the errors are seen as net biases for the individual interviewers and the effect is seen as the increase in variance due to the variability among these biases. The alternative approach is to consider the interviewer effect to arise from the creation of positive correlations among the response deviations contained in (almost all) survey data. Studies of interviewer variability date from the 1940s (see, for example, Mahalanobis 1946). The ANOVA model in this context was expounded by Kish (1962) and developed by Hartley and others; the correlation model was first presented by Hansen, Hurwitz and Bershad (1961) - the Census Bureau model - and developed by Fellegi (1964).

The other major component of imprecision in survey estimates is sampling variance. It is known that for most complex sample survey designs the precision of estimators is low compared to simple random sample designs of the same size. The loss of precision is due to the existence of positive correlations among characteristics for people belonging to the same area clusters, and area clusters typically form the sampling units for complex sample designs.

It is rare to find studies in which the complex sampling variance and the complex interviewer variance are both computed; Bailey, Moore and Bailar (1978) for the US National Crime Survey, and O'Muircheartaigh (1984a and b) for the World Fertility Survey in Lesotho and Peru are examples. This is due to a combination of design and analytic challenges. The norm for face-to-face interview surveys in both the US and UK is to have the workload from a given Primary Sampling Unit (PSU) assigned to a single interviewer and, moreover, to have each interviewer work in only one PSU. Such confounding is removed by an interpenetrated design in which respondents are assigned at random to interviewers. Due to cost considerations, these designs are rarely employed in face-to-face surveys. Even for telephone surveys, where the practical problems are less severe, though non-trivial (see Groves and Magilavy (1986)), such studies are rare.

This paper compares the relative impact of interviewer effects and sample design effects on survey precision by making use of an interpenetrated PSU/interviewer experiment which was designed by the authors for implementation in the second wave of the British Household Panel Study (BHPS). Section 2 of this paper describes in detail the data and methods used. Section 3 explores the results over all BHPS variables and illustrates on a few variables the use of a multilevel (hierarchical) approach in which the interviewer and sample design effects are estimated simultaneously while being incorporated in a substantive model of interest. Finally, Section 4 summarises and discusses our findings and their implications for survey research practice.

2. DATA AND METHODS

2.1 The BHPS and the Interpenetrated Design

The data source for this project is the British Household Panel Study (BHPS) which is conducted by the ESRC Centre for Micro-social Change at the University of Essex, UK. Interviewing on the BHPS began in 1991 and is scheduled to continue in annual waves until at least 1998. The survey used a multistage stratified cluster design covering all of Great Britain. The survey

[1] The data used in this paper were made available through the ESRC Data Archive. The data were originally collected by the ESRC Research Centre for Micro-Social Change.

instrument comprised a short household level questionnaire followed by a face-to-face 45 minute interview and short self-completion schedule with every adult in the household. Topics covered include household organisation, income and wealth, labour market experience, housing costs and conditions, health issues, consumption behaviour, education and training, socio-economic values, and marriage and fertility.

An interpenetrated design was implemented in a sample of PSUs in Wave II of the survey. Due to field requirements and travel costs, a constrained form of randomisation was adopted in which addresses were allocated to interviewers at random within geographic 'pools'; these pools are sets of 2 or 3 adjacent PSUs; twenty five 2-PSU pools, each with 2 interviewers, were used in the analyses which follow.

2.2 Analytic Methods

Our initial focus was on the calculation of intraclass correlation coefficients (ρ) for each of the components from the interpenetrated design. These included the interviewer (ρ_i) and the PSU (ρ_s). These coefficients were estimated for all variables in the dataset for which there were 700 or more responses. Categorical and most ordinal variables were transformed into binary variables prior to the analyses; ordinal attitude scales (Likert scales) were, however, treated as continuous. Hierarchical analyses of variance were then carried out for each of these variables using the SPSS MANOVA option. Data from the hierarchical analysis of variance runs were then assembled to create a meta dataset of ρ estimates. Other information was added to this dataset such as question type (attitudes, facts, quasi-facts, and interviewer checks) and topic area of the questionnaire.

2.3 Cross-Classified Multi-level Models

An alternative conceptualization of the analysis is as a multi-level (hierarchical) model in which the interviewer, PSU, and geographic pool are hierarchical partitions and the terms corresponding to them in the model are considered to be random effects. It is only recently that cross-classified multilevel analysis has become feasible (see Goldstein, 1995, Rasbash et al, 1995); the design is implemented in MLn by viewing one member of the cross-classification as an additional level above the other. A basic multilevel variance components model to capture the interviewer by PSU cross-classification within geographic pool can be defined as follows:

$$y_{i(jk)l} = \alpha + \beta x_{i(jk)l} + u_j + u_k + u_l + e_{i(jk)l} \qquad (1)$$

for the i th survey element, within the j th PSU crossed by the k th interviewer, within the l th geographic pool, where $y_{i(jk)l}$ is a function of an overall mean (α), an explanatory variable x and its associated coefficient β, and an individual error term ($e_{i(jk)l}$). Here u_j is a random departure due to PSU j, u_k is a random departure due to interviewer k, and u_l is the random departure due to geographic pool 1. Each of these terms and $e_{i(jk)l}$ are random quantities whose means are assumed to be equal to zero. In cases where the dependent variable is a dichotomy, $y_{i(jk)l}$ would be replaced by in (1) by log $(\pi_{i(jk)l}/1\text{-}\pi_{i(jk)l})$.

The treatment of the interviewer and PSU effects as random effects rather than as fixed effects (more common in the survey sampling literature) postulates a 'superpopulation' of interviewers from which the interviewers used in the study were drawn and an infinitely large population of PSUs. In the case of interviewers we can consider the inference as being made to the population of potential interviewers from whom the survey interviewers were drawn. For the PSUs the assumption involves essentially ignoring a small finite population correction. As we are interested in the relative magnitudes of the components of variance due to the interviewers and the sample design under the essential survey conditions this treatment will not affect our conclusions materially.

An added advantage of multilevel modelling in general, as recently demonstrated (cf Hox, de Leeuw, and Kreft, 1991; Wiggins, Longford and O'Muircheartaigh, 1992), is the facility to incorporate covariates directly into the analysis. For our work we will be able to examine such factors as interviewer age, gender, length of service, status, and whether the same interviewer was present for both Wave 1 and Wave 2 of the panel survey. We can also include respondent characteristics. We plan to add area level characteristics based on a match to census small area statistics in due course.

3. Results

3.1 Findings from Hierarchical Analysis of Variance

We present the results of this analysis in terms of the intraclass correlation coefficients for interviewers and PSUs. This coefficient measures the within-unit (cluster or PSU) homogeneity of the observations.

Within PSUs the homogeneity is a characteristic of the 'true values' of the elements in the population; the effect on the variance of an estimate is usually described using the *design effect* which is a function of the coefficient and the number of elements selected from within each PSU. The design effect is deff = 1 + ρ_s(b-1) where s denotes the sample clustering, ρ_s is the intra-cluster correlation, and b is the average cluster take. Within interviewer workloads the homogeneity results from the interaction between the interviewer and his/her respondents; the effect on the variance of an estimate may however be expressed in a form identical to that for the design effect. The *interviewer effect* is inteff = 1+ ρ_i(m-1) where i denotes the interviewer, ρ_i is the intra-interviewer correlation and m is the average interviewer workload. The cluster take and the interviewer workload arise as a result of decisions by the designer of the survey; ρ_s and ρ_i are quantities intrinsic to the population structure and to the quality of interviewers. As such the latter are more portable than the variance components themselves; the variance components themselves can of course be calculated once the ρ values are known.

During the past thirty years or so evidence has accumulated about the order of magnitude of both the intra-cluster correlation coefficient and the intra-interviewer correlation coefficient in sample surveys in the US and elsewhere. Though it is impossible to generalize with any confidence, the evidence suggests that values of ρ greater than 0.1 are uncommon and that positive values are almost universal for PSUs. In the case of ρ_i there is some evidence that by no means all variables are affected by interviewers in this way; attitude items and complex factual items are considered more sensitive to interviewer effect than simple factual items.

We included in the analysis 820 variables from the BHPS. Of these, 98 were attitude questions, 574 were factual, 88 were interviewer checks (items completed by the interviewers without a formal question), and 60 were quasi-facts (mostly on a self-completion form). Figures 1 and 2 below show the cumulative frequency distributions for ρ_s and ρ_i. The order of magnitude for the two coefficients were strikingly similar. As these values are themselves estimates they are subject to imprecision; using a test of significance at the 5% level 4 in 10 of the values of ρ_s and 3 in 10 of the values of ρ_i were significantly greater than zero. In the case of ρ_s this is not surprising as positive values are expected for most survey variables. What is somewhat surprising is that ρ_i is of the same order of magnitude, and that all types of questions seem to be affected. For attitude questions, 28% of the values of ρ_i were significantly greater than zero; for factual questions it was 24%; for interviewer checks, a staggering 59%; and for the quasi-factual self-completion questions, a more modest 17%. For these data, because of the way the investigation was designed, the average interviewer workload and the average cluster take were the same.

Figure 1: INTRA-INTERVIEWER CORRELATIONS

Cumulative Distribution of ρ_i

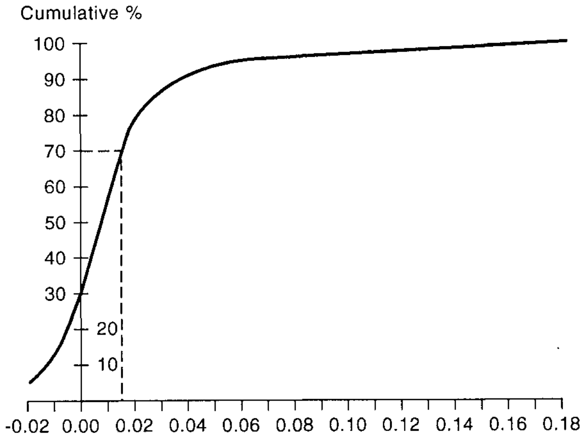
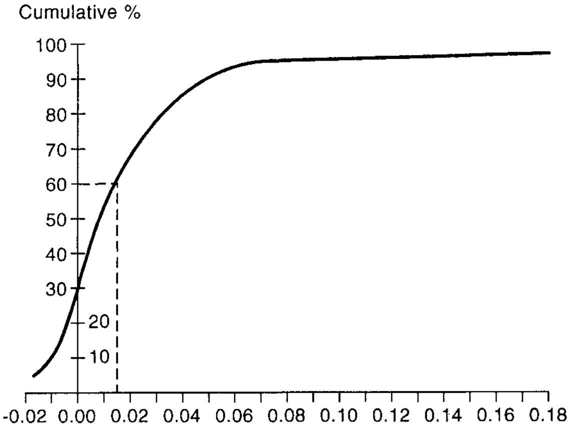


Figure 2: INTRA-CLUSTER CORRELATIONS

Cumulative Distribution of ρ_s



Furthermore, there was a clear positive correlation of 0.35 between ρ_i and ρ_s. Such a correlation has not, to our knowledge, been observed before. As variables are the elements in the computation of this correlation, the absence of such evidence may be because it is necessary to have a large number of variables to estimate such a correlation coefficient with any precision. In our analysis the correlation shows remarkable consistency across types of variables. The correspondence between ρ_s and ρ_i is by no means perfect; indeed it accounts for little more than a tenth of the variability in ρ_s and ρ_i.

It is not obvious what would bring about this correlation. A positive value implies that variables that show large intra-cluster homogeneity (show relatively substantial clustering among true values) are also sensitive to differential effects from interviewers. One possible explanation may be found in some of the early work on interviewers (see Hyman, 1954). Interviewer expectations are known to influence the responses obtained by interviewers. For a variable to have a relatively large value of ρ_s the individuals within a cluster will have relatively homogeneous values; it is possible that this consistency will affect the interviewers' expectations as the interviewer's workload progresses, leading to enhanced correlations within interviewer workloads.

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

3.2 Findings from Multilevel Models

A sample MLn model is shown in Table 1. This is one of the many interviewer check items which had large values of ρ_i. In this item, interviewers were asked to mark whether other people were present during the demographics section of the interview. The variable modelled is a binary subcategory indicating whether children were present. From the hierarchical analyses of variance, the estimated ρ values for this children present subcategory were $\rho_i = .171$ and $\rho_s = .062$. Model 1 is a basic variance components model showing the cross-classification of PSU and interviewer. Although the estimated standard errors of the random parameters are included in the table, the significance of the random parameters is based on a contrast test. We fund significant variation between interviewers but not between PSUs. In the model the estimate for variation between geographic pools was zero. [Parameters close to zero are often constrained to zero by the MLn programme; in this case the parameter remains zero even when employing the 'second order estimation procedure'.] In the standard formulation of the model the individual variation is assumed to have a binomial distribution and is constrained to 1.

In model 2, we have included the individual level explanatory variable, number of children in household, as it seems desirable to control for any systematic differences among interviewers in the composition of their workloads; an interviewer whose interviews take place in households without children would be expected to differ on this item from those interviewers whose workloads contained a large number of households with children. This control variable has a significant coefficient in the model. [For fixed effects significance may be judged by comparing the estimate with its standard error in the usual way.] It is interesting to note that the random coefficient for interviewer increases considerably. This suggests that it is not haphazard variation in interviewer workloads that explains this interviewer variability, but rather that the variation among interviewers in recording the presence of children is greater when opportunity (ie children in household) is taken into account.

We then proceeded to add in several interviewer explanatory variables. These included interviewer age, gender, status (whether basic interviewer, supervisor, or area manager), and years with the company. Also included was a measure of whether the same interviewer had visited the household for last year's interview. Of these various characteristics, only interviewer gender approached significance. Its addition to the model is shown under model 3. Here we can see that, though interviewer gender does contribute to the explanatory power of the model, the interviewer variance component is relatively unaffected.

There are at least two possible explanations for the correlated interviewer effect in this case. First, there is quite likely a difference in the ability of interviewers to arrange the circumstances of the interview so that the respondent is alone at the time - flexibility in making appointments, degree to which the interviewer emphasises the need for an undisturbed setting for the interview, etc. There is also the possibility that most of the between-interviewer variability is due to differences

Table 1: Multilevel Logistic Regression Model of IC Item: Children Present

	Model 1	Model 2	Model 3
Explanatory Variables	Fixed Effect (Std. error)	Fixed Effect (Std. error)	Fixed Effect (Std. error)
Grand mean	-1.05 (0.14)	-2.97 (0.28)	5.78 (1.53)
No. of children in Hh	-	1.55 (0.13)	1.56 (0.13)
Interviewer gender	-	-	1.47 (0.78)
Random Effects Source	Variance Component (Std. error)	Variance Component (Std. error)	Variance Component (Std. error)
Respondent	1	1	1
PSU	0.09 (0.12)	0.20 (0.24)	0.20 (0.24)
Interviewer	0.49 (0.20)	1.49 (0.49)	1.35 (0.47)
-2 log likelihood	786	730	499

in the extent to which, or the circumstances in which, interviewers record the presence of children; one source of variation could be in the definition of others being 'present'.

4. SUMMARY AND DISCUSSION

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

Our work with a specially designed study in wave II of the British Household Panel Survey (BHPS) permitted us to assess both these inflation components. Across eight hundred and twenty variables in the study, there was evidence of a significant impact of both the population clustering and the interviewers. The intraclass correlation coefficient, ρ, was used as the measure of homogeneity. We found that sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate. The median effect across the 820 variables was an 80% increase in the variance. We considered separately the different types of variables in the study, and found consistent effects across facts, attitudes, interviewer checks, and other items. The magnitude of the intra-interviewer correlation coefficients was comparable across these types, though the most sensitive items tended to be the interviewer check items. There was a tendency for variables which were subject to large design effects to be sensitive also to large interviewer effects and we offer a possible interpretation of this correlation in section 3.1.

We illustrate, using a binary interviewer check item - *children present during the interview*, a multilevel analysis (hierarchical modelling) that incorporates the sample design and interviewer effects directly into substantive models of interest. For this item we found a significant interviewer effect, our estimate of which was increased when we controlled for inequalities in the interviewers' workloads, and which persisted when we controlled for various extra-role characteristics of the interviewers. For other items not presented here we found situations where interviewer characteristics did help to explain the interviewer effects.

In later work we hope to explore further the factors which might provide an explanation of the variance components. From a modelling standpoint the issue is one of specifying appropriately the underlying factors in the substantive models of interest. From a sample survey standpoint the issue is that of incorporating in the analysis a recognition of the special features of the sample design and survey execution that make a particular data set deviate from *iid*. Multilevel analysis may provide a framework that makes it possible to reconcile the two approaches.

References

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

Fellegi, I. P. (1964), "Response Variance and Its Estimation", *Journal of the American Statistical Association*, 59, 1016-1041.

Goldstein, H. (1995). *Multilevel Statistical Models*(second edition). London: Edward Arnold; New York: Halstead Press.

Groves, R.M. and Magilavy, L.J. (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys", *Public Opinion Quarterly*, 50, 251-256.

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961), "Measurement Errors in Censuses and Surveys", *Bulletin of the International Statistical Institute*, 38, 2, 359-374.

Hox, J.J., de Leeuw, E.D., and Kreft, I.G.G. (1991), "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model", in *Measurement Errors in Surveys*, eds P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman. New York: John Wiley and Sons, Inc.

Hyman, H., (1954) *Interviewing in Social Research*, Chicago, University of Chicago Press.

Kish, L., (1962), "Studies of Interviewer Variance for Attitudinal Variables", *Journal of the American Statistical Association*, 57, 92-115.

Mahalanobis, P.C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute", *Journal of the Royal Statistical Society*, 109, 325-70.

O'Muircheartaigh, C.A. (1984a), "The Magnitude and Pattern of Response Variance in the Peru Fertility Survey", *WFS Scientific Report No. 45,* The Hague: International Statistical Institute.

O'Muircheartaigh, C.A. (1984b), "The Magnitude and Pattern of Response Variance in the Lesotho Fertility Survey", *WFS Scientific Report No. 70*, The Hague: International Statistical Institute.

Rasbash, J., Woodhouse, G., Goldstein, H., Yang, M., Howarth, J., and Plewis, I. (1995). *MLn Software.* London: Multilevel Models Project, Institute of Education, University of London.

Wiggins, R.D., Longford, N., and O'Muircheartaigh, C.A. (1992), "A Variance Components Approach to Interviewer Effects", in *Survey and Statistical Computing*, ed. A. Westlake, R. Banks, C. Payne, and T. Orchard. Amsterdam: North-Holland.

Woodhouse, G. (1995). *A Guide to MLn for New Users*. London: Multilevel Models Project, Institute of Education, University of London.