J. R. Statist. Soc. A (1998)
161, Part 1, pp. 63–77

The relative impact of interviewer effects and sample design effects on survey precision

Colm O'Muircheartaigh†
London School of Economics and Political Science, UK

and Pamela Campanelli
Social and Community Planning Research, London, UK

[Received May 1996. Revised February 1997]

Summary. One of the principal sources of error in data collected from structured face-to-face interviews is the interviewer. The other major component of imprecision in survey estimates is sampling variance. It is rare, however, to find studies in which the complex sampling variance and the complex interviewer variance are both computed. This paper compares the relative impact of interviewer effects and sample design effects on survey precision by making use of an interpenetrated primary sampling unit–interviewer experiment which was designed by the authors for implementation in the second wave of the British Household Panel Study as part of its scientific programme. It also illustrates the use of a multilevel (hierarchical) approach in which the interviewer and sample design effects are estimated simultaneously while being incorporated in a substantive model of interest.

Keywords: Interviewer effect; Interviewer variance; Multilevel models; Response variance

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 between the response deviations contained in (almost all) survey data; the increase in the variance of a mean is due to the positive covariance among these deviations. 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 Rao (1978) and others; the correlation model was first presented by Hansen *et al.* (1961) — the Census Bureau model — and extended by Fellegi (1964, 1974).

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 with that of simple random sample designs of the same size. Area clusters typically form the sampling units for complex sample designs and the loss of precision is due to positive correlations between people belonging to the same area clusters.

†*Address for correspondence*: Methodology Institute, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: colm@lse.ac.uk

0964–1998/98/161063

There are many other sources of measurement error in surveys. Some (e.g. coder variance) are relatively straightforward to estimate through either replication or interpenetration. Others (e.g. question wording effects) require special interventions in the survey process for their investigation. A comprehensive review may be found in Biemer *et al.* (1989).

Though there are some studies in which the complex sampling variance and the complex interviewer variance are both computed (Bailey *et al.* (1978) for the US National Crime Survey and O'Muircheartaigh (1984a,b) for the World Fertility Survey in Lesotho and Peru are examples), such studies are rare. This is due to a combination of design and analytic challenges. The norm for face-to-face interview surveys in both the USA and the 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. This confounds the sampling and non-sampling variances. Such confounding is removed by an interpenetrated design in which respondents are assigned at random to interviewers. Owing 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 uncommon.

Whereas it has been possible to carry out a simultaneous analysis of interviewer and cluster effects for sample means and other simple statistics, it is only recently that software has become available to estimate interviewer and cluster effects simultaneously while incorporating these effects directly into a substantive model of interest. This is possible through the use of a cross-classified multilevel model using the software package MLn (Rasbash *et al.*, 1995); alternative programs for multilevel analysis are VARCL (Longford, 1988) and HLM (Bryk *et al.*, 1986). (Note that, technically, means and proportions estimated from survey data are ratio estimates as there is uncontrolled variation in the sample size. For the British Household Panel Study (BHPS) the selection of PSUs with probability proportional to size and equal probabilities overall, this variation is fairly tightly controlled.)

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 BHPS. Section 2 describes in detail the data and methods used. Section 3 explores the results over all BHPS variables and illustrates the use of a multilevel (hierarchical) approach in which the interviewer and sample design effects are estimated simultaneously while being incorporated in a substantive model of interest. Finally, Section 4 summarizes and discusses our findings and their implications for survey research practice.

2. Data and methods

2.1. The British Household Panel Study and the Interpenetrated Design

The data source for this project is the BHPS which is conducted by the Economic and Social Research Council (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 wave 2 survey instrument comprised a short household level questionnaire followed by a face-to-face 45-minute interview and short self-completion schedule with every adult in the household. Topics covered included household organization, income and wealth, labour market experience, housing costs and conditions, health issues, consumption behaviour, education and training, socioeconomic values and marriage and fertility.

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

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

25 of the 30 usable geographic pools included two interviewers and two PSUs and five included three interviewers and three PSUs. Within PSUs in a given pool, households were randomly assigned to the interviewers working in those PSUs. The sample size for analysis of the 30 geographic pools was 1282 households and 2433 individual respondents.

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 data set for which there were 700 or more responses. (In general, the multivariate ANOVA (MANOVA) analyses which were used required 74 degrees of freedom. A rough rule of thumb to ensure sufficiently stable estimates is to set n greater than or equal to the degrees of freedom times 10. Applying this rule to the current models suggests an n of approximately 740.) Categorical and most ordinal variables were transformed into binary variables before the analyses; ordinal attitude scales (Likert scales) were, however, treated as continuous. Hierarchical ANOVAs were then carried out for each of these variables using the SPSS MANOVA option. The use of SPSS allowed us to explore this large number of variables more quickly and efficiently than would have been feasible with MLn. These hierarchical ANOVAs were restricted to cases from the 2×2 geographic pools as the program would not handle the simultaneous calculation of 2×2 and 3×3 geographic pools (note, however, that this is feasible with MLn). The elimination of the 3×3 geographic pools resulted in a reduction in sample size of 21% at the household level (to 1010 households) and 22% at the individual level (to 1903 individuals).

The sums of squares were partitioned using a 'regression approach' in which each term is corrected for every other term in the model. This makes sense substantively and also facilitates comparison with MLn. It also means that the values for ρ_i and ρ_s which are reported are conditional on each other. (As our design is not balanced, the sums of squares for the various components of the model will not add up to the total sum of squares. Also hierarchical ANOVA assumes a continuous dependent variable. For proportions between 0.20 and 0.80, however, the approximation should be fairly close.) Data from the hierarchical ANOVA runs were then assembled to create a meta data set of ρ-estimates constructed from the results of the 820 separate analyses of the original data. Other information was added to this data set such as question type (attitudes, facts, quasi-facts and interviewer checks) and topic area of the questionnaire.

2.3. Cross-classified multilevel models

An alternative conceptualization of the analysis is as a multilevel (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) and 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

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

for the ith survey element, within the jth PSU crossed by the kth interviewer, within the lth geographic pool, where $y_{i(jk)l}$ is a function of an appropriate constant α, explanatory variable(s) x and associated coefficients β, and an individual error term $e_{i(jk)l}$. Here u_j is a random departure due to PSU j, u_k is a random departure due to interviewer k, and u_l is the random departure due to geographic pool l. Each of these terms and $e_{i(jk)l}$ are random quantities whose means are assumed to be equal to 0. In cases where the dependent variable is a dichotomy, $y_{i(jk)l}$ would be replaced in equation (1) by $\log\{\pi_{i(jk)l}/(1-\pi_{i(jk)l})\}$, where

$$\pi_{i(jk)l} = \frac{\exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)}{1 + \exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)}.$$

When the dependent variable is continuous, ρ can be calculated directly from the variance estimates in a variance components model (e.g. interviewer variance divided by total variance). When the dependent variable is dichotomous, the variance components are given on the logistic scale and a more complex computation is required. We generate random normal deviates with variance given by the component estimate. These deviates are then transformed (taking the anti-logit) and the variance of these transformed values is calculated directly to give the numerator for ρ.

The treatment of the interviewer and PSU effects as random effects rather than as fixed effects (which is more common in the survey sampling literature) postulates a 'super-population' of interviewers from which the interviewers used in the study were drawn and an infinitely large population of PSUs. In the case of interviewers we can consider the inference as being made to the population of potential interviewers from whom the survey interviewers were drawn. For the PSUs the assumption involves essentially ignoring a small finite population correction (see, for example, Kalton (1979)). As we are interested in the relative magnitudes of the components of variance due to the interviewers and the sample design under the same essential survey conditions this treatment will not affect our conclusions materially.

An added advantage of multilevel modelling in general, as recently demonstrated (see Hox *et al.* (1991) and Wiggins *et al.* (1992)), is the facility to incorporate covariates directly into the analysis. For our work we are able to examine such factors as the age of the interviewer, 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 characteristics of the respondents. We plan to add area level characteristics based on a match to census small area statistics in due course. Single-level linear models have of course been used to analyse survey data. Such non-hierarchical models ignore the way in which the clustering in the sample design and the clustering of responses generated by the interviewers may affect the variance–covariance structure of the observations.

3. Results

3.1. Findings from hierarchical analysis of variance

The *design effect* is the most commonly used measure of the effect of within-PSU homogeneity on survey results; this is $\text{deff} = 1 + \rho_s(b - 1)$ where s denotes the clustering in the sampling frame, ρ_s is the intracluster correlation and b is the average number of elements selected from a cluster (the cluster take). We present the results of this analysis in terms of the intraclass correlation coefficients for interviewers and PSUs. Both measure the within-unit (interviewer or PSU) homogeneity of the observations. Within-PSU homogeneity is a characteristic of the true values of the elements in the population. Within interviewer workloads the homogeneity results from the interaction between the interviewer and his or her respondents; the effect on the variance of an estimate may, however, be expressed in a form that is identical with that for the design effect. The *interviewer effect* is $\text{inteff} = 1 + \rho_i(m - 1)$ where i denotes the interviewer, ρ_i is the intra-interviewer correlation and m is the average interviewer workload. The cluster take and the interviewer workload arise as a result of decisions by the designer of the survey; ρ_s and ρ_i are quantities that are 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 30 years or so evidence has accumulated about the order of magnitude of both the intracluster correlation coefficient and the intra-interviewer correlation coefficient in sample surveys in the USA and elsewhere. Though it is impossible to generalize with confidence, the evidence suggests that values of ρ_i greater than 0.1 are uncommon. (There is difficulty in comparing across studies as each involves different numbers of interviewers, different sample sizes and different types of variables. In addition, some researchers report the negative values of ρ_i which occur and others set these to 0.) Also, as indicated by the means in Table 1, the majority of values tend to be less than 0.02 (all these values are estimates, which accounts for the negative values in Table 1). There is also some evidence, although this is mixed, that different types of variables are affected by interviewers in different ways; attitude items and complex factual items are considered more sensitive to an interviewer effect than simple factual items are (see, for example, Collins and Butcher (1982), Feather (1973), Fellegi (1964), Gray (1956) and Hansen *et al.* (1961)).

The range of values reported in the literature for ρ_s is similar to that for ρ_i, though we would expect ρ_i to have more values near 0. Again, the evidence suggests that values greater than 0.1 are uncommon and that positive values are almost universal. The large values tend to be for certain types of demographic variables, notably tenure and ethnic origin. This is to be expected since adjacent groups of houses in a small area will tend to be of similar type and tenure, and people of similar ethnic origin often live close to each other (Lynn and Lievesley, 1991). Other demographic variables such as sex and marital status tend to show very low values. It is typically found that behavioural and attitudinal variables have ρ_s-values that are somewhere between these extremes, with attitudinal variables showing slightly lower values than behavioural variables. In the World Fertility Survey (see Verma *et al.* (1980)), the

Table 1. Summary of other interviewer variance investigations

Study	*Values of ρ_i*	*Mean*
Neighbour noise and illness (UK) (Gray, 1956)	−0.018 to 0.10†	0.015†
Television habits (UK) (Gales and Kendall, 1957)	(0.00) to 0.05, 0.19‡	§
Census (USA) (Hanson and Marks, 1958)	−0.00 to 0.061‡	0.011‡
Blue-collar workers (USA) (Kish, 1962)		
First study	−0.031 to 0.092	0.020
Second study: interview	−0.005 to 0.044	0.014
Second study: self-completion	−0.024 to 0.040	0.009
Census (Canada) (Fellegi, 1964)	(0.00) to 0.026	0.008
Health survey (Canada) (Feather, 1973)	−0.007 to 0.033	0.006
Mental retardation (USA) (Freeman and Butler, 1976)	−0.296 to 0.216	0.036
Aircraft noise (UK) (O'Muircheartaigh and Wiggins, 1981)	(0.00) to 0.09	0.020
Consumer attitude survey (UK) (Collins and Butcher, 1982)	−0.039 to 0.119	0.013
9 telephone surveys (USA) (Groves and Magilavy, 1986)	−0.042 to 0.171	0.009

†Calculated from *F*-ratios by using the formula supplied by Kish (1962).
‡Numbers available through Kish (1962).
§Mean cannot be computed: Gales and Kendall (1957) did not report all the variables analysed.

median ρ_s across various countries was 0.02 for various nuptiality and fertility variables. The median was much higher (around 0.08) for variables concerning contraceptive knowledge.

In comparing these two sources of variability, Hansen *et al.* (1961) found that the interviewer variance was often larger than the sampling variance. Bailey *et al.* (1978), in contrast, found response variance components that were at least 50% of their sampling variance for only a quarter of their statistics.

We included in the analysis 820 variables, some representing subcategories taken from BHPS items. 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). Fig. 1 shows the cumulative frequency distributions for ρ_s and ρ_i. The orders of magnitude for the two coefficients were strikingly similar. As these values are themselves estimates they are subject to imprecision; using a test of significance at the 5% level four in 10 of the values of ρ_s and three in 10 of the values of ρ_i

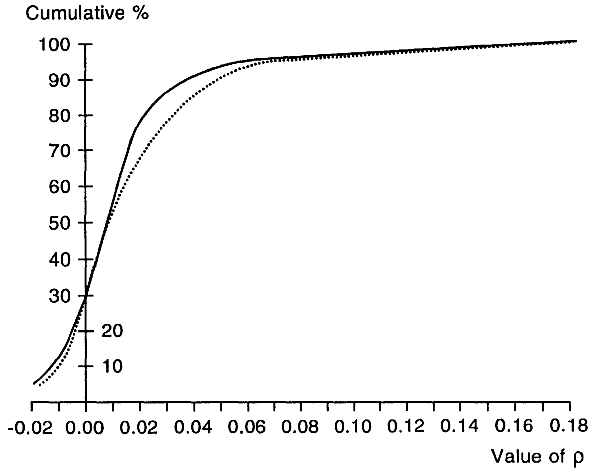

Fig. 1. Intra-interviewer and intracluster correlations: cumulative distribution of ρ_i (———) and ρ_s (··········)

were significantly greater than 0. In the case of ρ_s this is not surprising as positive values are expected for most survey variables. What is somewhat surprising is that, within the study, ρ_i is of the same order of magnitude. For these data, because of the way that the investigation was designed, the average interviewer workload and the average cluster take were the same; thus our estimates of ρ_s and ρ_i imply that the *effects* of the sample design and the interviewers were also about the same.

All types of questions show some significant values of ρ_i. For attitude questions, 28% of the values of ρ_i were significantly greater than 0; for factual questions it was 26%; for interviewer checks, a staggering 58%; for the quasi-factual questions, 25% (with the exclusion of the self-completion items). What is interesting is the similarity of the findings for the attitudinal and factual items, which is in contrast with the findings of some studies. There is some variation between types of attitudinal item. Among those items based on Likert scales, 33% showed significant values of ρ_i; this compares with 25% of the other attitude items.

We also looked for differences by source of the question. For example, 32% of the items in the individual schedule had ρ_i-values which were significantly greater than 0. The same was true for 17% of the self-completion items, 27% of the cover sheet items, 28% of the derived variables from the individual's questionnaire, 32% of the household questionnaire items and 34% of the derived variables from the household questionnaire. The notable difference here in susceptibility to interviewer effects is between the self-completion items and those that are interviewer administered. The fact that there is an interviewer effect at all on the self-completion form is interesting. Kish (1962), for example, found little evidence to suggest such an effect on the written questionnaires that he examined. O'Muircheartaigh and Wiggins (1981), however, did find an effect for a health supplement completed in the presence of the interviewer (as were the BHPS self-completion items).

There was also basically no difference in the proportion of significant ρ_i-values between the different sections of the questionnaire: demographics, health, marriage and fertility, employment, employment history, values and income and household allocation (with the percentage significant ranging from 22% to 35%). In contrast the section at the end of the questionnaire for interviewers to record their observations was highly susceptible to interviewer effects. 76% of the items in the interviewer observation section showed significant values of ρ_i. There was also a difference between dummy and continuous variables, with a higher proportion of effects being noted for the continuous variables.

Furthermore, there was a clear positive correlation of 0.35 between ρ_i and ρ_s. A positive correlation between ρ_s and ρ_i implies that variables that show large intracluster homogeneity (show relatively substantial clustering among true values) are also sensitive to differential effects from interviewers. Such a correlation has not, to our knowledge, been observed before. As the elements in the computation of this correlation are themselves variables, the absence of such evidence in the literature 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.

Homogeneous clusters contain individuals who are similar to one another; it is reasonable to suggest that individuals with *similar* values for the variable in question may respond in a *similar* way to whatever qualities the interviewer brings to bear in the interviewer–respondent interaction. This would mean that variables that manifested intracluster homogeneity would on balance be more likely than other variables to be display intra-interviewer homogeneity. An alternative explanation may be found in some of the early work on interviewers (see Hyman (1954)). Expectations of interviewers 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.

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

3.2. Findings from multilevel models

For illustration, we include three MLn models, one for each of the main types of variables: interviewer check items, facts and attitudes. These are shown in Tables 2–4 respectively. We have also shown the corresponding non-hierarchical (single-level) model to discover whether our substantive conclusions will be affected when we incorporate the data structure appropriately in the analysis.

The variable modelled in Table 2 is a binary subcategory indicating whether children were present during the demographics section of the interview, as noted by the interviewer. From the hierarchical analyses of variance, the estimated ρ-values for this *children present* subcategory were $\rho_i = 0.171$ and $\rho_s = 0.062$ ($n = 725$).

The hierarchical version of model 1 is a basic variance components model showing the cross-classification of PSU and interviewer. Although the estimated standard errors of the random parameters are included in Table 2, the significance of the random parameters is based on a contrast test. (This is recommended as the distribution of the standard errors for the random parameters may depart considerably from normality, especially in small

Table 2. Multilevel logistic regression model of the interviewer check item: children present†

	Model 1		*Model 2*		*Model 3*	
	Non-hierarchical	*Hierarchical*	*Non-hierarchical*	*Hierarchical*	*Non-hierarchical*	*Hierarchical*
Fixed effects						
Grand mean	−1.05	−1.05	−3.24	−3.30	−5.42	−5.49
	(0.08)	(0.14)	(0.37)	(0.41)	(0.94)	(1.27)
No. of children in household	—	—	1.20	1.23	1.23	1.25
			(0.10)	(0.11)	(0.10)	(0.11)
Respondent's gender (female)	—	—	0.62	0.59	0.62	0.60
			(0.21)	(0.22)	(0.21)	(0.22)
Interviewer's gender (female)	—	—	—	—	1.11	1.14
					(0.43)	(0.62)
			Variance components			
Random effects: source						
Respondent	—	1	—	1	—	1
PSU	—	0.09	—	0.08	—	0.08
		(0.12)		(0.17)		(0.17)
Interviewer	—	0.49	—	0.89	—	0.81
		(0.20)‡		(0.32)‡		(0.31)‡

†Standard errors are given in parentheses.

‡Significant random parameters based on a contrast test.

samples.) We found significant variation between interviewers but not between PSUs. In the model the estimate for variation between geographic pools for this variable was 0; this was not of course the case for all variables. Parameters close to 0 are often constrained to 0 by the MLn program; in this case the parameter remains 0 even when employing the 'second-order estimation procedure'. (In the estimation of random parameters in a logistic regression, MLn uses a weighted generalized least squares estimation procedure which requires the quantities to be estimated to be in the linear part of the model. A series expansion is used to approximate a linear form. Simulation and theory have suggested that the first-order estimation procedures can lead to an underestimation of the parameters. In many models the underestimation is negligible. However, in some models where predicted probabilities are extreme, or where there are few level 1 units per level 2 unit, underestimation can be severe. There is an option in MLn which allows the selection of a second-order estimation procedure. This procedure, however, is less computationally robust. See Woodhouse (1995) for a full description of this matter.) In the standard formulation of the model the individual variation is assumed to have a binomial distribution and is constrained to 1. (The validity of this assumption can be tested in MLn by relaxing this constraint.)

In model 2, we have included the individual level explanatory variable, number of children in household, as it is desirable to control for any systematic differences between interviewers in the composition of their workloads; an interviewer whose interviews take place in households without children would be expected to differ on this item from those interviewers whose workloads contained a large number of households with children. This control variable has a significant coefficient in the hierarchical model. (For fixed effects significance may be judged by comparing the estimate with its standard error in the usual way.)

Also included is the individual level explanatory variable 'respondent's gender'. We expected that the presence of children during the interview would be a function of the respondent's gender, with women respondents being more likely to have children with them than male respondents are. As can be seen by the values in Table 2, this expectation was confirmed.

It is interesting to note that the random coefficient for interviewers in the hierarchical version of model 2 increases in comparison with model 1. This suggests that it is not haphazard variation in interviewer workloads that explains this interviewer variability, but rather that the variation between interviewers in recording the presence of children is greater when opportunity (i.e. children in the household) is taken into account as well as the respondent's gender. The basic conclusion which can be drawn from model 2 is the same for both the hierarchical and the non-hierarchical versions of the model.

We then added several interviewer explanatory variables. These included interviewer age, gender, status (whether a 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 the previous year's interview. Of these various characteristics, only interviewer gender is considered in model 3. It was clearly significant in the non-hierarchical model and only approached significance in the hierarchical model. It is interesting that in this case different conclusions might have been reached depending on which model was considered. We also investigated the possibility of an interaction between interviewer gender and respondent gender. This coefficient was not significant under either version of model 3.

There are at least two possible explanations for the correlated interviewer effect in this case. First, it is quite likely that there is 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, the degree to which the interviewer emphasizes the need for an undisturbed setting for the interview, etc. There is also the possibility that most of the between-interviewer variability is due to differences in the extent to which, or the circumstances in which, interviewers record the presence of children; one source of variation could be in the definition of others being 'present'.

The key contrast here is between the message that we would obtain from ρ_i and ρ_s and the message from the multilevel analysis. With the former we would be concerned that the standard analysis would give spurious significance to the relationships estimated. In this case at least, however, an interviewer effect — though present for the dependent variable — does not affect the substantive analysis.

Table 3 deals with one of the respondent level factual items, newspaper readership. The variable modelled is a binary subcategory indicating whether or not the respondent typically reads the *Independent*. From the hierarchical ANOVAs, the estimated ρ-values for this readership subcategory were $\rho_i = 0.129$ and $\rho_s = 0.106$ ($n = 1268$).

Unlike the variance components model shown for the interviewer check item (see model 1), the basic variance components model given in model 4 shows a significant variation between PSUs as well as between interviewers. For this also there was no significant variation between geographic pools.

In model 5, we have included the individual level explanatory variable 'respondent's age'. Several other explanatory variables had also been explored in both the hierarchical and the non-hierarchical versions of the model (e.g. gender, social class, identification with a political party and income) but only respondent's age was significant. With this addition, the

Table 3. Multilevel logistic regression model of newspaper readership: whether the respondent reads the *Independent*†

	Model 4		*Model 5*		*Model 6*	
	Non-hierarchical	*Hierarchical*	*Non-hierarchical*	*Hierarchical*	*Non-hierarchical*	*Hierarchical*
Fixed effects						
Grand mean	−3.04	−2.99	−1.70	−1.94	−2.99	−3.19
	(0.13)	(0.30)	(0.35)	(0.45)	(0.67)	(0.90)
Respondent's age	—	—	−0.03	−0.03	−0.04	−0.03
			(0.01)	(0.01)	(0.01)	(0.01)
Whether same interviewer as previous year	—	—	—	—	0.21 (0.28)	0.63 (0.34)
Interviewer status						
Whether regular interviewer (compared with area manager)	—	—	—	—	1.35 (0.60)	1.06 (0.84)
Whether supervisor interviewer (compared with area manager)	—	—	—	—	2.25 (0.76)	2.23 (1.25)
			Variance components			
Random effects: source						
Respondent		1		1		1
PSU		1.55		1.48		1.59
		(0.64)‡		(0.63)‡		(0.66)‡
Interviewer		1.97		1.78		1.67
		(0.71)‡		(0.68)‡		(0.67)‡

†Standard errors are given in parentheses.

‡Significant random parameters based on a contrast test.

interviewer random variation is reduced slightly and the PSU random variation remains essentially the same.

Of the various interviewer explanatory variables we considered, two approached significance in the hierarchical version of model 6. These were the binary variable for whether the same interviewer had visited the household for the previous year's interview (interviewer continuity) and one of the two dummy variables modelling the three-category interviewer status variable (regular interviewer, supervisor or area manager). Here we can see that the interviewer variance component is again slightly reduced.

Interestingly we would have had a very different interpretation of which characteristics of the interviewer are having a significant effect if we had only run the non-hierarchical model. With the non-hierarchical model, the interviewer continuity variable was clearly not significant and the two interviewer status variables were clearly significant. In addition (although not shown in Table 3), the interviewer age variable approached significance. Middle-aged interviewers were more likely than elderly interviewers to record respondents as readers of the *Independent*.

Table 4 presents a behavioural intention item looking at whether or not the respondent expects to have any more children. As this is a subjective assessment, the question has been classified in the attitude category for our analysis. From the hierarchical analyses of variance, the estimated ρ-values for this item were $\rho_i = 0.075$ and $\rho_s = 0.048$ ($n = 1177$). As was the case for the variance components model under model 1, model 7 shows a significant variation between interviewers and possible variation between PSUs but not among geographic pools.

In model 8, we have included the three individual level explanatory variables number of children in the household, respondent's gender and respondent's age. Each of these is highly significant in both the hierarchical and the non-hierarchical versions of the model. With the

Table 4. Multilevel logistic regression model: whether the respondent is likely to have more children†

	Model 7		*Model 8*		*Model 9*	
	Non-hierarchical	*Hierarchical*	*Non-hierarchical*	*Hierarchical*	*Non-hierarchical*	*Hierarchical*
Fixed effects						
Grand mean	−0.39	−0.44	7.73	7.59	8.81	7.39
	(0.06)	(0.11)	(0.46)	(0.46)	(0.60)	(0.48)
No. of children in household	—	—	−0.85	−0.83	−0.86	−0.84
			(0.09)	(0.10)	(0.10)	(0.10)
Respondent's gender (female)	—	—	−0.65	−0.63	−0.64	−0.62
			(0.19)	(0.19)	(0.19)	(0.19)
Respondent's age	—	—	−0.24	−0.23	−0.24	−0.24
			(0.01)	(0.01)	(0.01)	(0.01)
Interviewer's years with company	—	—	—	—	0.042	0.043
					(0.020)	(0.027)
			Variance components			
Random effects: source						
Respondent	—	1	—	1	—	1
PSU	—	0.15	—	0.00	—	0.00
		(0.09)		(0.00)		(0.00)
Interviewer	—	0.22	—	0.38	—	0.34
		(0.10)‡		(0.16)‡		(0.15)‡

† Standard errors are given in parentheses.

‡ Significant random parameters based on a contrast test.

addition of these explanatory variables in the hierarchical model, random variation due to PSUs goes to 0 and random variation due to interviewers increases. The disappearance of the PSU effect may mean that the characteristics that led to the possible PSU effect have been adequately specified in the substantive model. Again, this suggests that it is not haphazard variation in interviewer workloads that is contributing to interviewer variability, but rather that there is variation between interviewers in their measurement of people's intentions to have more children.

In the non-hierarchical version of model 9, interviewer experience is a significant predictor with more experienced interviewers being more likely to record a 'yes' to the *more children* question than inexperienced interviewers are. Although not shown, in the non-hierarchical model, the interviewer continuity variable approached statistical significance. When the same interviewer returned on the second wave of the survey he or she was less likely to record yes to the more children question than a different interviewer was. These findings, however, do not hold for the hierarchical model.

Perhaps the most important point here is that, despite the strong interviewer effect, the substantive description represented by the substantive fixed part of the model is unaffected by the interviewers (at least not affected differentially). However, there are differences in the conclusions about the effect of interviewer characteristics depending on whether an interviewer variance term is explicitly included.

In addition to these examples, we conducted a further exploration of the effect of the extra-role characteristics of the interviewers (Sudman and Bradburn, 1974) on model conclusions. For each of the different types of item (attitudes, facts, quasi-facts and interviewer checks), a sample of variables was drawn from among those shown to have highly significant interviewer variability. Across the four categories, 26 items were drawn from 84. A cross-classified multilevel analysis (interviewer by PSU) was conducted on each of these with the interviewer characteristics as the explanatory variables. These included interviewer age, gender, status, years with the company and an indicator of interviewer continuity over time. Of the 26 models considered, interviewer age was significant in seven of the 26 cases (27%). The comparable percentages of significant effects that were found for the other interviewer characteristics were as follows: interviewer continuity, 12%; gender, 8%; interviewer status, 8%; years with the company, 4%. Although such data should be treated with caution, they may indicate that interviewer age is a general predictor of some of the interviewer variability on the high variability items. Freeman and Butler (1976), for example, found age and gender to be significant predictors of interviewer variance. Collins and Butcher (1982) also investigated the explanatory power of several characteristics of interviewers. Their strongest evidence was for an age effect.

Again we saw differences depending on whether a hierarchical or non-hierarchical model was used. The comparable figures for the non-hierarchical models were age significant in 27% of cases, interviewer continuity in 15%, gender in 12%, interviewer status in 35% and years with the company in 15%. In 11 of the 26 models, different conclusions about the effects of interviewer characteristics on substantive results would have been reached, depending on whether an interviewer variance term was explicitly included in the model.

4. Summarizing remarks 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 (e.g. inflation due to clustering) and execution (e.g. inflation due to correlated interviewer effects).

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

Our work with a specially designed study in wave 2 of the BHPS permitted us to assess both these inflation components. Across the 820 variables in the study, there was evidence of a significant effect of both the population clustering and the clustering of individuals in interviewer workloads. The intraclass correlation coefficient ρ was used as the measure of homogeneity. We found that sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate. The median effect across the 820 variables was an 80% increase in the variance. The magnitude of the intra-interviewer correlation coefficients was comparable across these types, though the most sensitive items tended to be the interviewer check items. There was a tendency for variables that were subject to large design effects to be sensitive also to large interviewer effects and we offer a possible interpretation of this correlation in Section 3.1.

The large values of ρ_i on particular items and the fact that ρ_i is of the same order of magnitude as ρ_s suggest that survey organizations should incorporate the measurement of ρ_i in their designs. If the necessary modifications of the survey design are too expensive to allow this, organizations should at least try to minimize its effect; this could be accomplished by reducing interviewers' workloads. Current practice tends to favour smaller dedicated interviewer forces with large assignments; in the presence of substantial interviewer effects this is a misguided policy.

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

Using software developed for multilevel analysis (hierarchical modelling) we have presented an alternative framework within which to consider the sample design and interviewer effects by incorporating them directly into substantive models of interest. For illustration we chose three binary items — an interviewer check item on *whether children were present during the interview*, a behavioural item, *readership of the Independent*, and a subjective item, *whether respondents thought it was likely that they would have another child.* For each of these items, we found a significant interviewer effect, which persisted when we controlled for inequalities in the interviewers' workloads and various extra-role characteristics of the interviewers. For other items not presented here we found situations where interviewer characteristics did help to explain the interviewer effects. In addition, we found that conclusions about the influence of the various extra-role characteristics would have differed in many cases if we had used only the standard non-hierarchical model rather than a hierarchical model.

In later work we hope to explore further the factors that might provide an explanation of the variance components. From a modelling standpoint the issue is 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 data. Multilevel models have a natural congruence with many important aspects of the survey situation; both the sample design and the fieldwork implementation can be described appropriately as introducing *hierarchical levels* into the data and thus multilevel analysis provides a framework that makes it possible to include both substantive and design factors in the same analysis.

Acknowledgements

The data were originally collected by the ESRC Research Centre for Micro-social Change at the University of Essex. The data file for this paper was made available through the ESRC Data Archive; the Archive bears no responsibility for the analyses or interpretations presented here.

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. *J. Am. Statist. Ass.*, **73**, 16–23.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N. and Sudman, S. (eds) (1989) *Measurement Errors in Surveys.* New York: Wiley.

Bryk, A. S., Raudenbush, S. W., Congdon, R. and Seltzer, M. (1986) *An Introduction to HLM: Computer Program and User's Guide.* Chicago: University of Chicago.

Collins, M. and Butcher, B. (1982) Interviewer and clustering effects in an attitude survey. *J. Markt Res. Soc.*, **25**, no. 1, 39–58.

Feather, J. (1973) A study of interviewer variance. *Report.* Department of Social and Preventive Medicine, University of Saskatchewan, Saskatoon.

Fellegi, I. P. (1964) Response variance and its estimation. *J. Am. Statist. Ass.*, **59**, 1016–1041.

——— (1974) An improved method of estimating the correlated response variance. *J. Am. Statist. Ass.*, **69**, 496–501.

Freeman, J. and Butler, E. W. (1976) Some sources of interviewer variance in surveys. *Publ. Opin. Q.*, **40**, 79–91.

Gales, K. and Kendall, M. G. (1957) An inquiry concerning interviewer variability (with discussion). *J. R. Statist. Soc.* A, **120**, 121–147.

Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd edn. London: Arnold.

Gray, P. G. (1956) Examples of interviewer variability taken from two sample surveys. *Appl. Statist.*, **5**, 73–85.

Groves, R. M. and Magilavy, L. J. (1986) Measuring and explaining interviewer effects in centralized telephone surveys. *Publ. Opin. Q.*, **50**, 251–256.

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961) Measurement errors in censuses and surveys. *Bull. Int. Statist. Inst.*, **38**, 359–374.

Hanson, R. H. and Marks, E. S. (1958) Influence of the interviewer on the accuracy of survey results. *J. Am. Statist. Ass.*, **53**, 635–655.

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

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: Wiley.

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

Kalton, G. (1979) Ultimate cluster sampling. *J. R. Statist. Soc.* A, **142**, 210–222.

Kish, L. (1962) Studies of interviewer variance for attitudinal variables. *J. Am. Statist. Ass.*, **57**, 92–115.

Longford, N. T. (1988) *VARCL Manual.* Princeton: Educational Testing Service.

Lynn, P. and Lievesley, D. (1991) *Drawing General Population Samples in Great Britain*. London: Social and Community Planning Research.

Mahalanobis, P. C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute. *J. R. Statist. Soc.*, **109**, 325–370.

O'Muircheartaigh, C. A. (1984a) The magnitude and pattern of response variance in the Peru Fertility Survey. *World Fertility Survey Scientific Report 45*. International Statistical Institute, the Hague.

———(1984b) The magnitude and pattern of response variance in the Lesotho Fertility Survey. *World Fertility Survey Scientific Report 70*. International Statistical Institute, the Hague.

O'Muircheartaigh, C. A. and Wiggins, R. D. (1981) The impact of interviewer variability in an epidemiological survey. *Psychol. Med.*, **11**, 817–824.

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

Sudman, S. and Bradburn, N. (1974) *Response Effects in Surveys*. Chicago: Aldine.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980) Sample designs and sampling errors for the World Fertility Survey (with discussion). *J. R. Statist. Soc.* A, **143**, 431–473.

Wiggins, R. D., Longford, N. and O'Muircheartaigh, C. A. (1992) A variance components approach to interviewer effects. In *Survey and Statistical Computing* (eds A. Westlake, R. Banks, C. Payne and T. Orchard). Amsterdam: North-Holland.

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