Survey Methodology, December 1995
Vol. 21, No. 2, pp. 99–106
Statistics Canada

The Effect of Interviewer Variance on Domain Comparisons

PETER DAVIS and ALASTAIR SCOTT[1]

ABSTRACT

In this paper we explore the effect of interviewer variability on the precision of estimated contrasts between domain means. In the first part we develop a correlated components of variance model to identify the factors that determine the size of the effect. This has implications for sample design and for interviewer training. In the second part we report on an empirical study using data from a large multi-stage survey on dental health. Gender of respondent and ethnic affiliation are used to establish two sets of domains for the comparisons. Overall interviewer and cluster effects make little difference to the variance of male/female comparisons, but there is noticeable increase in the variance of some contrasts between the two ethnic groupings used in this study. Indeed, the impact of interviewer effects for the ethnic comparision is two or three times higher than it is for gender contrasts. These findings have particular relevance for health surveys where it is common to use a small cadre of highly-trained interviewers.

KEY WORDS: Interviewer variance; Domain comparisons; Design effect.

1. INTRODUCTION

Surveys requiring a high degree of specialist training for interviewers, such as many health studies, are often forced to use a small number of highly-trained interviewers. There has been a substantial amount of work done on estimating the impact of interviewer variability on simple statistics such as means and proportions, and it is well-known that the use of a small number of interviewers, each having a high case load, can lead to a relatively large contribution to the total error. Comprehensive summaries of the literature are given in Groves (1989, chap. 8) and Lessler and Kalsbeek (1992, §11.3). However, most medical and social surveys are primarily interested in more complex questions such as comparisons between sub-groups or estimating the effect of a factor on disease outcome. There is a widespread belief that the effect of interviewer variability is much smaller here, and that the effect of a small number of interviewers is relatively harmless. Following the pioneering work of Kish and Frankel (1974), there has been a great deal of theoretical and empirical work on the effects of clustering on fitting multiple regression models or log-linear models for categorical data. Good accounts of the literature are given in Skinner *et al.* (1989) and Rao and Thomas (1988). There has been some empirical work on the conceptually simpler, yet practically important, problem of comparing sub-group means (see Kish 1987 and Skinner 1989 for example) but relatively little theoretical development.

In this paper we concentrate on comparisons between subgroups (or domains). We first look at theoretical aspects via a straightforward components of variance model. The theory suggests that the impact of interviewer variability depends on two things, the distribution of each interviewer's case load between the domains and the domain-interviewer interaction. Then we apply the theory to data from a reasonably typical health survey, using two sets of domains defined by the sex and ethnic background of the respondent. Unfortunately the study was not designed *a priori* to estimate interviewer effects (most importantly, interviewers were not deployed at random) so the results should be regarded as suggestive rather than definitive. However, they are sufficiently disturbing to indicate that the problem warrants further study. The results from the ethnic comparisons, in particular, suggest that there are cases when we should be concerned about using a small number of interviewers even when comparisons, rather than simple means or proportions, are the main concern of the analysis.

2. THEORY

For simplicity we start with the special case of a two-stage self-weighting design. This is sufficiently complex to illustrate the central ideas, but simple enough to avoid being swamped with extraneous detail. Following Collins and Butcher (1982), we want to address the problems of interviewer variance and clustering together. A simple correlated response model appropriate for observations drawn according to such a design is

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \qquad (1)$$

where i denotes the interviewer, p the primary sampling unit (PSU) and r the individual respondent. Here the

[1] Peter Davis, Department of Community Health, University of Auckland, Private Bag 92019, Auckland, New Zealand; and Professor Alastair Scott, Department of Mathematics and Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

mean, μ, is fixed constant and the remaining components, a_i, b_p and e_{ipr}, are assumed to be independent random variables with variances σ_I^2, σ_C^2 and σ^2 respectively. Such models have been used widely in theoretical studies of response variance. See Prasad and Rao (1990) for a recent example. For references to earlier work, see the comprehensive treatment in §11.3 of Lessler and Kalsbeek (1992).

Since the design is self-weighting the sample mean, $\bar{Y}$, is the natural estimator of the population mean. Its variance under the correlated response model (1) is

$$V(\bar{Y}) = (\bar{n}_I\sigma_I^2 + \bar{n}_C\sigma_C^2 + \sigma^2)/n$$

with $\bar{n}_I = \sum_i n_i^2/n$, where n_i is the number of respondents handled by the i-th interviewer and $n = \sum_i n_i$ is the total sample size, and $\bar{n}_C = \sum_p m_p^2/n$ where m_p denotes the number of respondents in the p-th PSU. Note that $\bar{n}_I$ is always larger than the simple arithmetic average of the n_i's and can be considerably larger if the n_i's vary widely.

Now consider what the corresponding expected variance, $V_0(\bar{Y})$ say, would be if the n observations had been generated independently (*e.g.* if we had drawn a simple random sample from a very large population of PSUs using a large pool of interviewers). It follows from (1) that

$$V_0 = \sigma_{t0t}^2/n \tag{2}$$

where

$$\sigma_{t0t}^2 = \sigma_I^2 + \sigma_C^2 + \sigma^2.$$

The inflation in the expected variance due to the combined effects of interviewer variability and intra-cluster correlation is given by the ratio

$$D_0 = V(\bar{Y})/V_0$$
$$= 1 + (\bar{n}_I - 1)\rho_I + (\bar{n}_C - 1)\rho_C \tag{3}$$

where $\rho_I = \sigma_I^2/\sigma_{t0t}^2$ and $\rho_C = \sigma_C^2/\sigma_{t0t}^2$. We shall refer to this ratio as the "design effect" although it differs slightly from the usual definition which is in terms of actual, rather than expected, variances. It is clear from expression (3) that interviewer variability can have a substantial effect on the variance of a sample mean if the average interviewer case-load, $\bar{n}_I$, is large even if the intra-interviewer correlation, ρ_I, is relatively small.

Next suppose that we are interested in the difference between two domain means rather than a single mean. We might, for example, be interested in gender differences or in differences between two ethnic groups. In the simplest extension of the correlated response model (1) we might postulate a model of the form

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i + b_p + e_{ipr}^{(d)} \tag{4}$$

for observations from the d-th domain. Here the means, $\mu^{(d)}$, may be different for the two domains but the interviewer and cluster effects are assumed to be the same.

Let $p_i^{(d)} = n_i^{(d)}/n^{(d)}$, where $n_i^{(d)}$ is the number of respondents from domain d contacted by the i-th interviewer and $n^{(d)}$ is the total number of respondents from domain d. Similarly, let $q_p^{(d)} = m_p^{(d)}/n^{(d)}$, where $m_p^{(d)}$ is the number of respondents from domain d lying in the p-th PSU. Then, under model (4), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$, the difference between the sample means for the two domains, is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) =$$
$$(\bar{m}_I\sigma_I^2 + \bar{m}_C\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right), \tag{5}$$

where

$$\bar{m}_I = \sum_i (p_i^{(a)} - p_i^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \tag{6}$$

and

$$\bar{m}_C = \sum_p (q_p^{(a)} - q_p^{(b)})^2 \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right). \tag{7}$$

If the observations had been generated independently the corresponding expected variance would be

$$V_1 = \sigma_{t0t}^2\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

so that the inflation due to interviewer variability and intra-cluster correlation is now

$$D_1 = \mathrm{Var}(\bar{Y}^{(a)} - \bar{Y}^{(b)})/V_1$$
$$= 1 + (\bar{m}_I - 1)\rho_I + (\bar{m}_C - 1)\rho_C. \tag{8}$$

The size of the effect depends on the way the interviewers' case-loads and the PSUs cut across the domains. At one extreme, when each interviewer contacts the same proportion of people from both domains, (*i.e.* when $p_i^{(a)} = p_i^{(b)}$), $\bar{m}_I$ is zero and the interviewer effect essentially cancels out. At the other extreme, when each interviewer sees only cases from a single domain, $\bar{m}_I$ is similar in size to $\bar{n}_I$ and the interviewer effect for differences is comparable to that for a single mean. Typically interviewers contact people from both domains and $\bar{m}_I$ is rather small, giving some justification to the belief that interviewer variability has a small impact on estimated differences between domains. Similar comments apply to the effect of clustering.

All this depends on the assumption that the interviewer and cluster effects, a_i and b_p, are the same for both domains. It is easy to imagine situations where such an assumption would not be at all reasonable. Some interviewers, for example, might interact very differently with males and females, or with members of different ethnic groups. A model which allows for the possibility of such interactions is

$$Y_{ipr}^{(d)} = \mu^{(d)} + a_i^{(d)} + b_p^{(d)} + e_{ipr}^{(d)}, \qquad (9)$$

where $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are now assumed to be correlated random variables with correlation $r_I(r_C)$. The naive model (4) corresponds to the special case in which the variances of the effects are equal and r_I and r_C are both equal to one. On the other hand, if there are substantial differences between the interviewer (cluster) effects for the two domains, $r_I(r_C)$ will be small (or even negative in extreme cases). In the rest of this section we suppose for simplicity that the variances of $a_i^{(a)}$ and $a_i^{(b)}$ (respectively $b_p^{(a)}$ and $b_p^{(b)}$) are equal. This may or may not be reasonable in practice but the simplification enables us to concentrate on the essential ideas. The basic form is similar in the more general case but the terms are somewhat messier. Under model (9), the expected variance of $\bar{Y}^{(a)} - \bar{Y}^{(b)}$ is

$$V(\bar{Y}^{(a)} - \bar{Y}^{(b)}) = (\bar{v}_I\sigma_I^2 + \bar{v}_c\sigma_C^2 + \sigma^2)\left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right) \qquad (10)$$

where

$$\bar{v}_I = \sum_i \left(p_i^{(a)2} - 2r_I p_i^{(a)}p_i^{(b)} + p_i^{(b)2}\right) \Big/ \left(\frac{1}{n^{(a)}} + \frac{1}{n^{(b)}}\right)$$

with a similar definition for $\bar{v}_C$ in terms of $q_p^{(a)}$ and $q_p^{(b)}$.

The variance inflation factor under this model is

$$D_2 = 1 + (\bar{v}_I - 1)\rho_I + (\bar{v}_C - 1)\rho_C. \qquad (11)$$

This is a decreasing function of r_I, the correlation between the interviewer effects for the two domains; the smaller the correlation, the larger the variance inflation. When $r_I = 1$, $\bar{v}_I$ reduces to $\bar{m}_I$ and the interviewer effect is negligible provided all interviewers see a reasonable balance of people from both domains. However, if r_I is small (indicating a strong interaction between the interviewers and domains), $\bar{v}_I$ is the same order of magnitude as $\bar{n}_I$ and the effect of interviewer variability on the variance of domain differences can be substantial.

In practice, the effects will fall between the two extremes and their likely impact is a matter for empirical enquiry. In the next section, therefore, we make a start on building up practical knowledge about the impact using data for a variety of questions drawn from a single health survey that is typical of the genre of research investigation for which domain comparisons are important (although not ideally designed for our purposes!).

3. EXAMPLE

The example is based on data drawn from a survey of the oral health, attitudes and practices of adult New Zealanders. The details of the survey are reported in full elsewhere (Cutress *et al.* 1979). The important features of the study for the purposes of the current investigation are the sample design and the deployment of interviewers.

The sample design was a stratified multi-stage sampling scheme. The country was divided into 256 Territorial Local Authorities (TLAs) and a geographically stratified sample of 68 TLAs was drawn from the 256 with selection probabilities proportional to size (PPS) at the first stage, where size was the estimated number of persons aged 15 and over. Each sampled TLA was split into secondary sampling units (SSUs) comprising existing census mesh-blocks, aggregated where necessary in order to achieve a minimum size of 50. Two SSUs were then selected with PPS from each sampled TLA at the second stage. Finally, a systematic sample of 28 adults was drawn from each sampled SSU. This equalised the final probability of selection for all adults so that the sample design is (approximately) self-weighting.

The key point of the design was the deployment of the interviewers. Thirteen interviewers were employed in the study, with at least three interviewers used within each SSU, and all interviewers carried out at least 10% of their total work-load in one region (Auckland). Ideally the assignment of interviewers would be part of the overall sample design as in Fellegi (1974) or Biemer and Stokes (1985). Unfortunately the study was not designed to estimate interviewer variance, and the assignment of respondents to interviewers was done in a haphazard way, rather than using a formal randomization procedure.

This is fairly typical of large studies. The following quote from Hox (1994) gives a good summary of the situation: "Ideally, in interviewer studies, respondents should be assigned to interviewers at random. In large-scale studies, this is seldom done because it is expensive and complicated to organize. This makes it difficult to use such studies for methodological research because, as a result, interviewer and respondent characteristics might be confounded. Multi-level analysis, as outlined above, offers some remedies for this situation. If the relevant respondent variables are known they can be put in the regression model to equalize interviewers by statistical means... The limitation of this approach is that it relies on statistical control instead of experimental control. It depends on the assumption that all relevant covariates have been included

and have been correctly modeled. Without randomization, it is impossible to conclude that the influence of all confounding variables has been eliminated." In our case, the deployment is such that all the components of variance are formally identifiable, provided that we believe the model and are willing to accept that the assignment of interviewers is independent of the cluster effects. However, because of the lack of formal randomization there always remains the possibility that variations in patterns of response between interviewers could be a function of workload allocation rather than interviewing style. Clearly the empirical results can only be regarded as tentative, pointing out possibilities that will need to be explored further in properly designed studies.

Even if we ignore the lack of randomization in the interviewer deployment, the study design is considerably more complicated than the one assumed for the development of the theory in the previous section, since it involves three stages of sampling and regional stratification of the first stage units. In the full analysis, we fitted a more complex model including fixed effects terms for the stratification, a hierarchical random effects model for the three stages of sampling, and all second-order interaction terms. However it turned out that the TLAs used as the first stage units were so diffuse that the differences between strata and the between-TLA component of variance were negligible for all the variables used in the following analysis. Thus the between-SSU component is dominant and, for all practical purposes, we can treat the design as if it were a two-stage sample with the meshblocks (aggregated where appropriate) as PSUs. We have ignored the other components in the results reported in the next section.

4. RESULTS

We look first at interviewer and cluster effects on a selection of means and proportions. We have used Model (9) for both types of variable. It is now well-known that this leads to an under-estimate of the variance components for binary data (see Anderson and Aitken 1985 and Pannekock 1988 for example), so our estimated design effects for proportions should be regarded as lower bounds. The models are fitted using PROC GLM in SAS. The impact of clustering has been well documented in the literature (Kish 1965; Kish and Frankel 1970; 1974). In general terms, the magnitude of the effects of clustering depends on the type and number of units chosen and is likely to vary with different kinds of social and demographic characteristics. In the current investigation clustering effects were expected to be reasonably high because the census meshblocks used as sampling units are likely to show a fair degree of internal homogeneity. In keeping with this concentration of population characteristics, it was assumed that demographic and related items would show the largest values of ρ_C. Values of ρ_I were expected to be lower because of the intense interviewer training. The literature suggests that these effects are also likely to vary according to the type of questionnaire item, with attitude questions, questions requiring probing, fixed-alternative and forced-choice items, together with poorly-worded and ambiguous questions, being particularly susceptible to interviewer variability (Feather 1973, Groves 1989).

Estimated measures of intra-interviewer and intra-cluster correlation coefficients for a selection of questionnaire items falling under four separate headings (socio-demographic, attitudinal, reports of recent behaviour, and recall of distant behaviour) are outlined in the first two columns of Table 1. These categories were identified as providing natural groupings with the potential to display a wide range of interviewer effects. Within each grouping the items are listed in order of the size of their intra-interviewer correlations. A full description of all questionnaire items (apart from the self-evident socio-demographic category) is provided in the Appendix.

As expected, the socio-demographic variables (except for gender) show the highest values of intra-cluster correlation. The average ρ_C is .07 (.08 if gender is omitted). The average values of ρ_C for the other three categories of item are .02 and less. A few items that might be expected to be closely related to social background - like dental visiting, payment for visits, toothbrushing and certain attitude statements - have higher than average ρ_C values. In general, though, these values fall within the range reported by others. (See, for example, Kish 1965, p. 581 for a series of consumer surveys, Bebbington and Smith 1977 and Verma *et al.* 1984 for the country studies in the World Fertility Survey.)

The corresponding estimated ρ_I values are listed in the first column of Table 1. In general these values are very much smaller than those recorded for cluster effects, being usually less than half, and in some cases a tenth, the size of the ρ_C values for the corresponding items. As expected, some attitude items show higher than average ρ_I values, as do certain reports of behaviour that might be susceptible to a high "social desirability" bias, like toothbrushing and buying sweets and chocolates. Ethnic group and employment status also record relatively high values. The pattern is similar to that found in previous studies, although the values recorded lie at the lower end of the range of typical values reported elsewhere (Feather 1973; Kish 1962; O'Muircheartaigh 1977; O'Muircheartaigh and Wiggins 1981). A comprehensive survey is given in Chapter 8 of Groves (1989). This may partly reflect the intensive training and monitoring of the interviewers that were integral to the field work stage of the study. It may also be influenced by the rigorous post-field work "cleaning" (editing and checking) of the data that was carried out prior to analysis. However it may also simply be due to the attenuation resulting from using Model (1) for proportions that we noted above.

Table 1

Cluster and Interviewer Effects for Means and Proportions

Item Description	$\hat{\rho}_I$	$\hat{\rho}_C$	D_0	%Int
Attitudinal:				
Dentists 1	.014	.014	4.61	91
Visiting	.008	.028	3.42	74
Natural Teeth	.008	.027	3.52	76
Health of Teeth	.007	.015	2.97	84
Dentures	.005	.015	2.67	80
Dentists 2	.004	.033	2.77	57
Health of Gums	.003	.010	1.96	77
Fluoridation	.001	.016	1.66	49
Average	**.006**	**.020**	**2.95**	**73**
Socio-demographic:				
Employment Status	.010	.055	4.20	65
Race	.009	.172	6.87	33
Age	.004	.042	5.98	52
Household Income	.002	.092	3.29	15
Marital Status	.000	.058	2.34	0
Sex of Respondent	.000	.005	1.12	0
Average	**.004**	**.071**	**3.47**	**28**
Recent Behaviour:				
Brushed Teeth	.019	.025	6.16	8
Sweets/Chocolates	.011	.003	3.75	98
Fluoride Toothpaste	.008	.000	3.04	100
Toothpick	.006	.006	2.66	92
Rinse Mouth	.004	.024	2.43	62
Dental Floss	.001	.018	1.60	43
Disclosing Tablet	.000	.027	1.49	0
Mouthwash	.000	.018	1.42	0
Average	**.006**	**.012**	**2.82**	**60**
Distant Behaviour:				
Age First Paid	.004	.029	2.34	57
Visited Dentist	.004	.029	2.51	57
Cost Last Year	.002	.000	1.19	100
Year Last Visit	.000	.014	1.15	0
Average	**.003**	**.018**	**1.80**	**54**

Perhaps more significant than the pattern and values of ρ_I is the impact of interviewer variability on the overall design effect, incorporating both interviewer and clustering effects. This is shown in the third column of Table 1 (D_0), with the final column (% Int) representing the proportionate contribution of interviewer variability to the overall value of D_0. Design effects are substantial, being above two in all but a minority of cases. This is due to the clustering and to the impact of the large interviewer workloads characteristic of the study since, from equation (3), the variance is increased by a factor of $1 + (\bar{n}_I - 1)\rho_I$, where $\bar{n}_I$ is a weighted average of the interviewer workloads. There is a distinct pattern in the contribution to the design effect produced by interviewer variability. For socio-demographic variables it averages just under one half of the contribution from clustering, while for attitudinal items the interviewer contribution to the design effect rises to three times that from clustering. The other two categories of items range in between these two extremes.

What the results outlined in Table 1 confirm is the impact that interviewer workload has on the variance of sample estimates, because of the multiplier effect. In essence, an interviewer component with a very small intra-class correlation can be translated into a major effect if the interviewer workload is high. In the study under review, the logistics of deployment and the requirements of on-going quality control seemed to argue for small interview teams, a practice that appears to be typical of much field work in the health area (for example, Choi and Comstock 1975). This meant that interview workloads averaged over 250. The cost of this strategy is immediately apparent from the results in Table 1; very small differences between interviewers are translated into major reductions in the precision of sample estimates.

Now we turn to the main object of our analysis, viz. the impact of interviewer variability on contrasts between domain means or proportions. In the current analysis, this was assessed for two sets of comparisons, the first set by gender (male/female) and the second set by ethnic group (European/non-European). As we have seen in the discussion following equation (11), the contribution, $1 + (v_I - 1)\rho_I$, to D_0 from interviewer differences depends on the extent to which the interviewer effect is constant across the two domains and on the way the domains cut across individual case-loads. Assuming that the domains cut evenly across interviewer case-loads, then v_I is zero if the interviewer effect is identical in the two domains, in which case the common interviewer effect cancels out completely in the comparison. On the other hand, if the effects in the two domains are weakly correlated then the value of v_I tends to be much higher and in extreme cases may equal the average case-load. In the current study values of v_I fell between 0 and 50 for both gender and ethnic group. Thus the effect of domain-specific interviewer effects on the design effect can be quite substantial. Similar comments apply to the impact of clustering on the comparison; if the effect is the same on both domains then it largely cancels out and the net impact is small, but the impact can be substantial if the clustering effect is domain specific.

Table 2 shows values of ρ_I and ρ_C for comparisons by gender and by ethnic group, together with the overall design effect D_2 and the proportion of this effect due to the impact of interviewer variability. Note that the item on the use of disclosing tablets has been omitted from Table 2. This is because so few respondents either used or knew what this item was that the effective sample size in this case is tiny, thus rendering the results almost meaningless.

The impact of both interviewer and clustering on comparisons by gender is small with design effects little above unity, in spite of the fact that the estimated values of ρ_I and ρ_C are slightly increased when adjusted for this variable.

Table 2

Interviewer and Cluster Effects for Domain Differences

Item Description	By Sex				By Race			
	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int	$\hat{\rho}_I$	$\hat{\rho}_C$	D_2	%Int
Attitudinal:								
Dentists 2	.004	.043	1.05	0	.010	.027	1.28	46
Visiting	.009	.028	1.08	0	.072	.133	5.19	78
Natural Teeth	.010	.032	1.06	0	.010	.037	1.26	42
Fluoridation	.001	.019	1.12	42	.021	.031	2.13	88
Dentures	.007	.018	1.04	0	.011	.035	1.21	33
Health of Teeth	.012	.022	1.52	85	.010	.045	1.40	20
Dentists 1	.001	.018	1.05	0	.015	.020	1.53	74
Health of Gums	.006	.022	1.07	0	.003	.104	1.46	9
Average	**.006**	**.025**	**1.12**	**16**	**.019**	**.054**	**1.93**	**49**
Socio-demographic:								
Race	.008	.183	1.11	0	-	-	-	
Household Income	.004	.095	1.37	24	.004	.099	1.95	40
Marital Status	.000	.059	1.17	0	.011	.060	1.69	38
Employment Status	.014	.067	1.42	71	.022	.116	2.09	25
Age	.007	.052	1.06	0	.006	.093	1.87	24
Sex of Respondent	-	-	-		.006	.011	1.09	44
Average	**.007**	**.091**	**1.23**	**19**	**.010**	**.076**	**1.74**	**34**
Recent Behaviour:								
Brushed Teeth	.025	.060	1.62	65	.019	.019	1.68	88
Rinse Mouth	.007	.029	1.28	64	.004	.023	1.20	45
Mouthwash	.000	.057	1.20	0	.027	.105	2.69	75
Dental Floss	.003	.021	1.06	0	.015	.036	1.37	32
Toothpick	.006	.010	1.03	0	.006	.046	1.48	63
Sweets/Chocolates	.012	.009	1.02	0	.013	.022	1.31	48
Fluoride Toothpaste	.010	.007	1.11	100	.007	.000	1.02	100
Average	**.009**	**.028**	**1.19**	**33**	**.013**	**.036**	**1.36**	**64**
Distant Behaviour:								
Age First Paid	.003	.033	1.10	0	.029	.141	2.92	71
Visited Dentist	.005	.035	1.04	0	.020	.018	1.26	50
Year Last Visit	.004	.012	1.20	75	.016	.003	1.83	12
Cost Last Year	.007	.021	1.01	0	.076	.117	2.09	42
Average	**.005**	**.025**	**1.09**	**19**	**.035**	**.070**	**2.03**	**44**

A significant gender-specific effect was apparent for only three items, health of teeth and tooth-brushing - for which there may be a unique social acceptability bias - and employment status - which holds quite different connotations for men and women. Note that the interviewer effect is the dominant one in all three of these comparisons.

The impact on comparisions by ethnic group is much higher, with design effects averaging about 1.7. This suggests that there are significant, non-cancelling interviewer and clustering effects associated with the ethnic identity of respondents. There are large ethnic-specific interviewer effects for two hypothetical attitudinal questions (visiting and fluoridation), for one item of recent behaviour, and for age of first payment for dental services. The result is plausible; all the interviewers were European and may have varied systematically in their interactions with respondents of different ethnic backgrounds. Again clustering effects are most marked for the socio-demographic variables. Not only are the design effects on average higher than those recorded for the gender comparisons, but the interviewer component is in general two or three times higher for the ethnic group contrasts.

A referee rightly points out that because of the way the interviewers are deployed (they worked primarily in teams assigned to different parts of New Zealand), there is a real possibility that the interviewer effects might be inflated because of confounding with area effects. The fact that differences between the TLAs were so small gives us some reason to believe that this inflation will be small, but the possibility can never be discounted with this design.

5. DISCUSSION

This paper has applied empirical data from a not untypical health survey to assess the impact of interviewer variability under the assumptions of both simple and extended versions of the correlated response model for the error variance of a multi-stage sample design.

In the first case the simple model analyses the relative impact of cluster and interviewer effects on the estimation of means and proportions. The results of this analysis confirm a number of findings that are well established in the literature: the intra-class correlations for interviewers are generally lower than those for clusters; the intra-class correlations for clusters vary in the expected direction by question type; the overall design effects for these question types vary between 2 and 3.5; a substantial component of this inflation is contributed by interviewer variability and can probably be attributed to the multiplier effect of large interviewer caseloads; finally, the impact of this interviewer component is shown to vary in the expected direction by question type.

In the second case the extended model addresses the analysis of cluster and interviewer effects for the estimation of domain contrasts between means and proportions for two sets of comparisons defined by gender and ethnic group. The effect on contrasts between domain means was smaller but it was still significant for a number of items, particularly for the ethnic comparisons, suggesting that the interviewer effect was different for the two domains. The size of the effect for these items was certainly large enough to suggest that we should be concerned about it in designing similar studies. In general, the impact of interviewer effects was two or three times as great for the ethnic contrasts as it was for the gender comparisons.

The basic deficiencies in the design mean that these results must be regarded as suggestive rather than definitive. They do indicate, however, that there is considerable potential for damage in the use of a small group of interviewers even when interest is centered on domain differences rather than simple means or proportions. This is certainly counter to standard folklore in some fields such as health surveys, and suggests that considerable further empirical work is justified.

On the assumptions of the simple correlated response model a reduction in the impact of interviewer variance

can be achieved by raising the number of interviewers and thus reducing individual interviewer workloads. Of course, this brings with it a potential reduction in the quality of interviewing if training and monitoring procedures have to be tempered. In this instance close attention to question wording and interviewer instruction is clearly crucial. In the case of the extended version of the correlated response model, however, such a strategy is unlikely to be a sufficient one on its own. If comparisions between groups are a major objective of the study, then it is important also to ensure that the interviewers treat the two groups in as similar a way as possible. It is also important to design the study so that each interviewer contacts respondents drawn from both groups. This is likely to be a critical consideration in investigations such as case-control studies in which health outcomes are related to contrasting exposures and in which the control of potential confounder variables may have a significant influence on the magnitude of measures such as the odds ratio.

ACKNOWLEDGEMENTS

This project has received support from the Medical Research Council of New Zealand. The bulk of the detailed computations for this paper were carried out by Joanna Broad.

APPENDIX
Questionnaire Items

Attitudinal

Dentists 1: "Dentists are more interested in their patients than making money."

Dentists 2: "Dentists recommend a lot more things to be done than really need to be done."

Dentures: "Dentures are just as good (or better) than your own teeth."

Fluoridation: "What is your opinion on fluoridating public water supplies?"

Visiting: "Do you think a person should go to the dentist only when they have dental problems or should they go sometimes also when they have no obvious problems?"

Health of Teeth: "If you went to the dentist tomorrow, do you think he would find anything wrong with your teeth?"

Health of Gums: "If you went to the dentist tomorrow do you think he would find anything wrong with your gums?"

Recent Behaviour

"Yesterday did you
- use a dislosing tablet/mouthwash/dental floss/toothpick?
- rinse after eating?
- brush your teeth?"

"Did you buy sweets or chocolates any time last week?"

Distant Behaviour

Age First Paid: "About how old were you when you first went to a dentist for routine treatment for which you or your family had to pay?"

Visited Dentist: "Did you visit a dentist in the last 12 months?"

Year Last Visit: "In what year did you last visit a dentist?"

Cost Last Year: "About how much did you pay for dental treatment in the last 12 months?"

REFERENCES

ANDERSON, D.A. (1985). Variance component models with binary response; interviewer variability. *Journal of the Royal Statistical Society*, Series B, 47, 203-210.

BIEMER, P.B., and STOKES, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 158-166.

BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis, *The Analysis of Survey Data*, (C.A. O'Muircheartaigh and C. Payne, Eds.), 175-192. New York: John Wiley.

CHOI, I.C., and COMSTOCK, G.W. (1975). Interviewer effects on responses to a questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.

COLLINS, M., and BUTCHER, B. (1992). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

CUTRESS, T.W., HUNTER, P.B., DAVIS, P.B., BECK, D.J., and CROXSON, L.J. (1979). *Adult Oral Health and Attitudes to Dentistry in New Zealand*, Medical Research Council, Wellington.

DIJKSTRA, W. (Ed.) (1982). *Response Behaviour in the Survey Interview*. New York: Academic Press.

FEATHER, J. (1973). *A Study of Interviewer Variance*, (WHO/ICS-MCU Saskatchewan Study Area Reports Series 2, No. 3). Department of Social and Preventive Medicine. University of Saskatchewan, Saskatoon.

FELLEGI, I.P. (1974). An improved method of estimating correlated response variance. *Journal of the American Statistical Association*, 69, 496-501.

GROVES, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

GROVES, R., and FULTZ, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research*, 14, 31-52.

HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, 143, 474-487.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Society*, 57, 92-115.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

KISH, L., and FRANKEL, M.R. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley.

LESSLER, J.T., and KALSBEEK, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley.

O'MUIRCHEARTAIGH, C.A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, 10, 97-115.

O'MUIRCHEARTAIGH, C.A., and PAYNE, C. (Eds.) (1977). *The Analysis of Survey Data*, (Volume 2: Model Fitting). New York: John Wiley.

O'MUIRCHEARTAIGH, C.A., and WIGGINS, R.D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine*, 11, 817-824.

PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.

PRASAD, N.G., and RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classified data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.

VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C.A. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Series A, 143, 431-473.