

Cluster Sampling: A False Economy?

Andrew Zelin and Roger Stubbs
MORI

BACKGROUND AND CONTEXT

The random sample survey is a commonly used method in market research and such surveys generally are considered to provide more robust results than purposive or convenience (such as quota) samples. Simple random samples (SRS), where every member of the population has an equal or predetermined chance of selection, and the sampling is single stage, are the simplest to visualise and the level of precision of results derived from them is easy to calculate. However, SRS are almost impossible to achieve in reality (due to imperfect sampling frames, non-response, and so on) and, in any case, they are relatively inefficient ways of obtaining a particular number of responses, in terms of fieldwork and travelling costs.

Much of the theory underlying this paper is not new, and many of the concepts covered relate to important work published a number of years ago. However, the messages that this paper serves to portray are no less relevant today than they were in the past, as cluster sampling continues to be a popularly used type of survey design. It is less statistically efficient than simple random sampling, and this is as critical as ever today as researchers continue their quest to obtain data in an ever-faster and less expensive way.

If one considers the situation of running a national survey of 10,000 people across the whole of Great Britain, it would be impractical and expensive to send interviewers across the whole country to addresses which are completely randomly scattered, because travelling time and interviewer costs are likely to be prohibitive. It would be more economical and realistic to focus on specific localities (or clusters) across the country (e.g. constituencies or wards) and sample from some localities (these particular localities being selected at random) but not from others. There is, however, a price to pay for this saving in administrative time and costs, namely that one would receive no information about the localities or the clusters from which one is not sampling. Correspondingly, aggregate results would be based only on the clusters from which one is sampling. Also, there may be good reasons why people's views within a small area are similar. This would lead to a reduction in precision and a widening of confidence intervals as compared to the situation of running a survey on a totally SRS basis.

This paper begins by providing a brief background to the concept of the design effect, and how it can be used to see how various aspects of survey design (such as clustering, but also stratification and weighting) can impact on the level of statistical reliability of a survey. Focusing on clustering, a commonly used formula for calculating the confidence interval (CI) for clustered samples is then introduced. Using this formula on a worked example covering a specific national survey, it is shown how the level of confidence can vary with the number of clusters taken (which is controllable) and with the differences between the clusters (which cannot be controlled). At this point, what appears to be a surprising conclusion is verified by an alternative methodological approach. The paper then goes on to present a theoretical and somewhat historical example which nevertheless is useful in illustrating that the design effect is very different for different survey measures. The paper also touches on the issue of interviewer variability, which is in itself a form and extension of clustering. Finally, the report is put together with appropriate costings to look at the practical and financial implications of clustering-related losses of effective sample size.

The Concept of the Design Effect

When calculating the variance or the CI of a sample, based on a percentage measure (e.g. the percentage of residents who are satisfied with refuse collection in their area, or who vote Labour), the standard simple random sample (binomial) formula is typically assumed, i.e.: (See [Equation 1](#))

where p is the proportion of respondents (of a sample of n), expressing a particular opinion.

In practice, samples are never totally of the simple random type. Clustering, weighting, finite populations and stratification can make the true variance different from that of the theoretical one as calculated by [Equation 1](#). The extent to which it differs – that is, the ratio of the true to the theoretical variance – is known as the design effect (DE). With finite populations and stratification (assuming it is appropriately applied), the DE is less than 1.0, thus leading to a narrowing of CI (or an improvement in precision). With weighting and clustering (in most cases), the opposite occurs. The DE is thus the extent to which the statistical reliability is made better or worse than a similar-

sized SRS and can be defined as the ratio of the actual sample size to the effective sample size. (See [Equation 2](#))

With clustering, the lower the proportion of clusters which are sampled of all the clusters in the survey area, the greater the DE (and the lower the precision). Also (in terms of the measure of interest, e.g. the percentage of people who are satisfied with a particular service) the greater the difference *between* the clusters that are being sampled then the greater the difference is likely to be among *all the clusters* in the survey area, and hence the lower the precision and the greater the DE will be above 1.00.

Ideally, clusters should be microcosms of the whole population and be internally heterogeneous (i.e. have the full range of variability within them) and be externally homogeneous (i.e. be as similar to each other) as possible. The greater the extent to which this is fulfilled, the closer the DE would be to 1.00 and so the lower the statistical price paid for reducing travelling costs and time.

The most important factors which drive the DE of clustered samples are how many clusters are sampled and how different the results are from one cluster to another. With the second issue, this is very sensitive to the question that is asked, because with some questions there may be great differences between clusters, while for others the clusters might be virtually identical. *Design effect can differ greatly from question to question with a particular survey (and even between response categories at a given question), and this is very much the case with clustered sampling.*

CALCULATING THE DE AND RELIABILITY OF CLUSTERED SAMPLES

The most popular, traditional textbooks on sampling theory (e.g. Hansen *et al.* 1953; Kish 1965; Cochran 1968) provide thorough algebraic details on how to calculate DEs and statistical reliability for clustered samples. However, and despite the popularity of the use of clustered samples, relatively little exists in the literature on how this work can be used by market researchers in practice. One useful and relevant paper by Paul Harris, (1977) which was republished in the JMRS in 1997, uses the concept of intra-class correlation to assess the DE of clustered samples. The intra-class correlation coefficient (which in this paper will also be referred to as the intra-cluster correlation) ρ is a measure of similarity of units within clusters. Alternatively, using Cochran's formula (described below), one can calculate and summate the two components of variance in a clustered sample – that is, the *within* and the *between* components. As with Harris's (1997) work, this paper seeks to take the textbook descriptions a stage further by looking at measures of trade-off (in terms of accuracy) between the number of clusters used and the number of units sampled *within* clusters, and hence the trade-off between precision and cost.

So how can the DE, or at least the variance of a sample result, be calculated for a clustered sample? Generally, it is necessary to calculate the standard error (SE) for clustered samples, and then calculate the SE for the equivalent SRS of the same overall sample size. The DE is the ratio of the squares of the SEs (i.e. the ratio of the variances).

The DE is generally: (See [Equation 3](#))

The formula for the variance of a clustered sample (i.e. SE^2_{cl}) around which this paper is based is given below and one should note that this formula makes the assumption that the sizes of the clusters are equal (Cochran, 1968, p. 279): (See [Equation 4](#))

Where n is the number of clusters in the sample, m is the (mean) number of respondents in each cluster, f_1 is the proportion of clusters in the population which are sampled ($=n/N$; N being the total number of clusters in the population), f_2 is the proportion of respondents in each cluster who are sampled, p_i = percentage of respondents who show a characteristic *within* cluster i , is the mean percentage of respondents who show a characteristic and $q_i = (1 - p_i)$.

The central issue is that the variance of a clustered sample has two components, which add together. These are:

- The *between*-cluster component (to the left of the plus sign).
- The *within*-cluster component (to the right of the plus sign).

One would need to know how much difference there is in the results between different clusters. Imagine that one is conducting a council amenities satisfaction survey across different local authorities. Many studies carried out by MORI have shown that respondents *within* the same wards would have more similar levels of satisfaction than *between* those in different wards. This might be due to similar demographics or outlooks, but more importantly because tenants in the same wards are likely to have the same ultimate service provider and experiences of that service provider. As will be shown later, when one conducts sensitivity tests, the DE depends very heavily on the extent of the *between*-cluster variability. It is virtually impossible to say what this between-cluster variability is likely to be until one has actually conducted the study and collected the results. Therefore, without information on the cluster-to-cluster differences, attempting to come up with an accurate statistical reliability figure for a survey

which involves a clustered sample can be little more than a 'finger-in-the-wind' exercise. Thus here lies the potential value of ρ – the intra-cluster correlation (discussed in more depth later) – and possibly DE figures, which is 'portable' from one survey to another. So, if one knows something about the pattern in an earlier survey, one is at least able to make sensible forecasts.

WORKED EXAMPLE: NATIONAL SATISFACTION STUDY ACROSS LOCAL AUTHORITIES

To put this into context, an example of a national satisfaction survey conducted on 10,000 respondents across 30 geographical units is illustrated here. Those respondents *within* each unit were taken to represent all adults within their respective unit and the 30 LAs covered were taken to represent 450 nationally. Each respondent was asked the extent to which he/she was satisfied with their home overall. Those who responded as a 1 or a 2 on the original 5-point scale were considered (for the purpose of this exercise) as 'satisfied', with everyone else being 'not satisfied'. This gave a simple 'yes or no' (or 'Binary') measure to use, which could be aggregated up at LA or national level as a 'percentage of respondents who are satisfied' measure.

The DE due to the clustering is calculated, firstly by working out the variance of the estimate of the national 'percentage satisfied' as a clustered sample. This is then recalculated as if it were a simple random, unclustered sample, and the ratio of the two variance figures gives the DE (i.e. the ratio of the *clustered* : *un-clustered* variance). To obtain the clustered-sample variance, one needs to know how different the clusters are in terms of their percentage satisfied. This critical 'cluster-to-cluster distribution' information is available from the actual survey results, aggregated at unit level – see the Appendix.

A common question asked by market researchers at the design or proposal stage of projects is what the size of the DE due to clustering is likely to be. This is a very difficult question to answer as it can vary enormously between studies and between questions on the same study and is affected by all the raw ingredients of the mathematical formulae. These include, among other things, the number of clusters in the sample and the population, the number of residents and respondents in the survey within each cluster and the *between*-cluster variability in the results. DEs due to clustering can potentially exceed 20, or in extreme (but mostly rare and theoretical) circumstances, be less than 1; in other words, make a sample more precise than the equivalent SRS. The approach taken involves looking at how sensitive the DE is to each of the different measures in the survey – that is, a sensitivity analysis. It could help answer such questions as:

- How many extra respondents would be needed to compensate for there being a big difference *between* the clusters, given that sampling a proportion of clusters may be unreliable if the difference *between* the clusters is large?
- If one has a budget for 10,000 interviews, how much better is it to run 200 interviews in each of 50 areas as compared to 500 interviews in each of 20 areas? How many extra interviews is the difference equivalent to in terms of precision? What is the financial benefit and does this pay for the extra travelling costs of having to visit more parts of the country?

Varying the Number of Clusters Used

The starting-point for this analysis will involve taking 30 geographical units (local authorities) from a population of 450 nationally. The total sample size is 10,000 respondents, so *within* each unit there is a sample of 333 (= an average of 10,000/30), of a population of 45.7 million adults nationally (and therefore an average of 101,600 in each LA). The proportion of clusters nationally which are included in the survey (f_1) is 30/450 or 0.067. The proportion of the population in each selected cluster who are interviewed (f_2) is 0.003 (i.e. 333/101,600). The sum, across the 30 units, of the squared differences in proportions who are satisfied, from the overall proportion (79.5% – see the Appendix) is 0.105 while the sum of the component is 4.63.

Had this have come from a simple, non-clustered sample, the statistical reliability would be $\pm 0.8\%$; in other words, the CI on a finding of 79.5% would be $\pm 0.8\%$ (or 78.7%, 80.3%), as calculated using the standard binomial formula below. (See [Equation 5](#))

Putting all of the information described in the previous paragraph into the cluster sample formula, one obtains a statistical reliability which is considerably wider (i.e. a DE of 6.96, or an effective sample size which is reduced from 10,000 to 1437; dividing the actual sample size by the DE gives the effective sample size) – see [Figure 1](#). The 30-cluster situation yields a CI of $\pm 2.1\%$ (or 77.4%, 81.6%).

Without changing the total actual sample size, but increasing the sample coverage from 30 to 40 units (and therefore 250 respondents per unit), the DE falls markedly to 5.1 and the effective sample goes up to 2000. With 50, 60 and 80 units, the effective sample increases respectively to 2500, 3100 and 4200. Indeed, even with 100 clusters (using this dataset), one still does not actually see a benefit of clustering (over an SRS). Although rare in practice, it is theoretically possible for a clustered sample to give tighter CIs than an SRS, where the clusters themselves provide a greater level of coverage/representation of areas than would an SRS. This may be because with clustering, one would be controlling for a certain (minimal) amount of coverage of respondents within the

selected clusters — something not guaranteed in simple random sampling. It should be noted also that the point at which one starts to see a benefit depends on how similar or different the clusters are to each other in terms of the results they yield, as described shortly. At the extreme end of the spectrum, when all of the clusters are sampled, one effectively has a *stratified sample*, and these are typically associated with increased levels of precision over SRS. Harris (1997) states that negative values of ρ occur where the clusters are more uniform than would be produced by random sampling and these situations can occur for certain characteristics, such as gender and age. [Figure 2](#) shows the effect of increasing the number of clusters covered on the DE. It can be seen that precision diminishes very rapidly at a low number of clusters, but at the other end of the scale, it is very much a law of diminishing returns.

With a different amount of *between*-cluster variation, the position of the curve and the break-even point (where the DE = 1) would be different. However, the overall shape and potency of the relationship between the number of clusters and the statistical reliability would be largely unaffected.

Going back to the question about trading off (20 clusters \times 500 units per cluster vs 50 clusters \times 200 units per cluster), keeping the sample size the same, one would obtain effective sample sizes of 920 and 2500 respectively. This means that by visiting 30 extra clusters, one obtains 1580 extra effective interviews in terms of precision, or increasing the sample nearly threefold. In financial terms, this is likely to be a bargain price to pay for all of these extra effective interviews. If one were to insist that the number of clusters could not be increased beyond 20 (despite sampling a virtually unlimited number of units *within* these 20 clusters), then it would be very difficult to increase the effective sample size or reduce the width of the CI beyond a negligible amount. In fact, even increasing the total sample from 10,000 (i.e. 20 clusters \times 500) to 100,000 (20 \times 5000) keeps the CI at $\pm 2.60\%$, which represents a very poor use of resources.

Researchers may be interested in looking to see whether DEs can be improved by altering the geography of the clusters and this is described more fully later in the paper. But let us look at what would be likely to happen if the LAs were replaced by wards, which are smaller? The overriding factor would be whether or not the number of clusters would increase. In other words, whether, say, an LA cluster would be replaced by three separate ward units (increasing the number of clusters and maintaining, to some extent, the proportion of the included population), or whether it would be replaced just by a single ward (i.e. keeping the *number* of cluster units used unchanged). Increasing the number of cluster units would naturally reduce the DE (see below) as the predominant effect. However, this is likely to be offset by the fact that any differences between survey measures for small wards are likely to be greater than differences between larger constituencies. An LA, being larger than a ward, is itself the more likely of the two to represent the overall national population. It would also be necessary to take account of the effects of the proportion of the population which fall into sampled clusters (i.e. f_1 from Cochran's formula), and the proportion of respondents in the population within a cluster which is covered in the survey (i.e. f_2). Thus such experimentation would be complex and form the basis of a significant amount of further useful research.

The Concept of Intra-cluster Correlation (ρ)

At this point, it may be worth introducing an alternative mathematical approach to the area of design effects and cluster-to-cluster differences in the survey measures. Work done by Collins and Goodhardt (1978) applies the intra-cluster correlation approach and verifies the assertion of the limited benefit of large cluster sizes in contributing to the levels of statistical precision.

The intra-cluster correlation measure ρ is a measure of how much of the total respondent-to-respondent variability in a survey measure is accounted for by differences *between* clusters. If, for example, one is running a survey of pupils across a number of schools (with the schools being the clusters), then ρ is a measure of the extent to which the opinions of pupils *within* the same school resemble each other as compared with those from children at different schools (Rasbash *et al.* 2000). The ρ measure is potentially more useful than the DE measure as it drills down to the level of detail of a specific survey measure, and as will be seen later, can vary substantially from question to question within the same survey. However, unlike the DE, it does not take account of the general survey design information, such as the number of the clusters and their size.

The DE is related to ρ and is calculated by the following formula (where m is the average cluster size): (See [Equation 6](#))

The assertion that the narrowness of the confidence interval of a survey estimate can be limited, regardless of the number of respondents in a cluster (presented towards the end of the previous subsection) may, on the face of it, seem surprising. However, [Table 1](#) shows how dramatically the DE increases with the cluster size for a fixed amount of intra-cluster correlation/similarity. The effect is even more apparent when the information is expressed in terms of the effective sample size per cluster (i.e. m/DE) – see [Table 2](#).

This powerful table shows that no matter how many respondents there are in a cluster, there is only a finite or optimal level of precision or effective sample size which can be achieved. This effective sample size is the reciprocal of the intra-cluster correlation figure.

It can therefore be seen intuitively how the number of larger units (or clusters) taken is so much more critical to the reliability than the number of respondents taken per larger unit. However, why might this be so?

Looking at Cochran's formula discussed earlier, the variance (and hence the statistical reliability) of a clustered sample has two key components which are additive: a *within*-cluster component and a *between*-cluster component. If one were carrying out a clustered survey of 1000 respondents and were to move from a design of 50 respondents in 20 clusters to 25 respondents in 40 clusters, the *between* and the *within* components would change in size, but in opposite ways. The *within* component would double in size (in changing from 50×25 to 25×40) but the *between* component would reduce to half of what it was. The overall effect on the total variance effectively depends on which of the two components is the larger in absolute terms. Unless the cluster size (m) is very small and the proportion of clusters in the population that are covered is large, then the dominant term in the majority of cases would be the *between* component. Therefore increasing the number of clusters covered would virtually always reduce the sample variance and increase the effective sample size. This is typically close to a proportional relationship whereby a doubling of the number of clusters covered leads to a doubling of the effective sample size.

Varying the Between-cluster Difference or Spread of Results

The other big determinant of statistical reliability/DE of clustered samples involves *how different the clusters* are from each other in terms of the results they yield. The ideal situation is where each cluster is a microcosm of the entire population and contains the full amount of variability which the population contains. Pure examples of attributes/measures which are obviously correlated with clusters include aspects such as age and gender (see later), but also, for example, medically-related measures. (For instance each cluster in a survey sample might be expected to contain people with high, low and middling red blood cell counts.) Here, the aggregate cluster results/means would be very similar to each other. In these situations, the statistical reliability can exceed those of simple random samples of the same size (thus yielding DEs of less than 1.0). At the other extreme, people within a cluster may have very similar survey characteristics to each other, but different from those within different clusters. Here, not only is one less likely to capture the full sample variation when using a clustered design, but there is likely to be a great deal of uncertainty about what results the non-sampled clusters would have yielded.

One can demonstrate this in the satisfaction survey illustrated. In the table given in the Appendix, the middle column shows the percentage of respondents who are satisfied with local services in each unit (or cluster) and they range from 87.3% to 62.9%. Using this amount of variation and a total sample of 10,000 spread across 30 clusters, one obtains a DE of 6.96. However, if, for example, the spread of cluster results were doubled (see right-hand column of the Appendix), cluster 1 would be twice as far away from the national figure as it was before. This doubling of the level of spread of the clusters practically quarters the effective sample size. On the other hand, halving the cluster-cluster variation has the opposite effect. With the example in [Figure 2](#), the situation involving 30 clusters and 333 responses in each cluster gives a DE of 7.0. If one doubles the spread of the cluster measures about the same overall national average of 79.5%, then the DE increases approximately fourfold to 28. Halving the spread leads to a four-factor reduction (i.e. to 1.8).

Therefore, just as the *number of clusters* is having a powerful effect on the DE, the *cluster-cluster difference* in results is also having a powerful effect, as shown in [Figure 3](#). Mathematically, this is attributable to the term in the larger, more influential, *between*-cluster part of the formula in A1.

The Combined Effect

Finally, putting both components (i.e. the number of clusters and the spread of results) together, it can all be summarised as in [Figure 4](#) overleaf. As mentioned earlier, regardless of the cluster-to-cluster spread of the results, there is still a strong relationship between the number of clusters and the statistical reliability.

In practical terms, Harris (1977) recommends 'the selection of natural clusters of sampling units which are as heterogeneous as possible, as then less efficiency is lost by cluster sampling'. This would be a wise policy and would be dependent on one knowing, before running the survey, which clusters within one's sampling frame fall into this category. Indeed, the whole issue of being able to estimate and minimise design effect due to clustering is very difficult without prior knowledge of what the results (and the levels of difference between the clusters) are likely to be, before one conducts the fieldwork. The data which 'feed' Cochran's formulae are only available after collection of the results. As a guide, it may be possible to estimate this type of information prior to running the survey by looking at results of similar surveys conducted previously, although one would need to bear in mind that the DE would vary from question to question within a survey. This is why it is helpful to build a 'library' of p values, which is question/topic-based rather than survey-based. In principle, researchers could tap into this to establish likely values for the key variables in any survey.

EXAMPLE: HOW THE DE CAN DIFFER BETWEEN QUESTIONS IN THE SAME SURVEY

Cochran (1968, pp. 66–68) gives a very vivid example of how dramatically the DE can change from one question in a survey to another. Consider an attempt to measure the following in a household survey: (1) the sex ratio; and (2) the percentage of people who visited a doctor in the last 12 months.

This took place within an area covering 15,000 households, of which 30 households ('households' in this case can be considered as the clusters) were sampled. For each, Cochran worked out these measures, along with their statistical reliabilities, based on (1) a clustered design and (2) for comparison, an SRS design. He found that with visiting the doctor, there was a DE in excess of 1.00, while the opposite occurred with measuring the sex ratio. The reason for this is that household (cluster) members are likely to see their doctors in similar patterns to each other. It is likely that if one member has a tendency to seek medical advice readily for even minor symptoms, then they all are likely to ('if one goes, they all go'), and vice versa. Consequently, one is likely to see quite a wide range in the percentage of household members who have seen their doctor, and in quite a few households, all have (100%) while in others, none have (0%). This high cluster-to-cluster variation explains the high design effect.

With the sex ratio, the opposite occurs, as most households are centred quite closely at 50:50, with few all-male or all-female households. After all, the typical nuclear family would contain at least one of each sex. This low level of cluster-to-cluster variation explains the fact that better reliability can be obtained from such a clustered sample than the equivalent random sample. (Please note at this point that this is a study which relates to 1947 Census data in Baltimore, USA. Unlike today, there were very few single-person households, because people did not live so long (so there were fewer single pensioners), divorce was less common and young people tended to remain in their parental home until they married. Nevertheless, this historical example has been included as it clearly illustrates the contrast between the DEs based on the two measures, namely sex ratio and doctor visits.)

INTERVIEWER VARIABILITY

Clusters that arise in sample designs may not always reflect geographic distinctions, as in the examples highlighted above. Even if a simple random (or stratified) sample is being carried out with complete effective geographic coverage, it is unlikely that a single interviewer would be covering the entire sample (in which situation concerns about overall bias would arise). There would naturally be a number of interviewers working on the project, and it is understandable that results collected from one interviewer would be more similar to each other than those from different ones. This may be down to the different ways in which they ask questions, the impression they make on respondents and ways in which the responses are interpreted and recorded. Alongside geographic clustering, different interviewers tend to be allocated to specific clusters, and this effectively enhances and exacerbates any geographical clustering effects. Collins noted that: 'p can represent the contribution of interviewers to total variance', and he showed (1983) that interviewer variability can change from one type of survey question to another. On a survey of consumer attitudes, the question of whether or not the current time is perceived as a good time to spend evoked an intra-interviewer correlation (ρ) of 11.5% with opinions about the country generally attracting the highest such figures. Opinions about themselves (respondents) had a ρ of 1.0%, while factual questions had the lowest ρ overall (0.7%).

It would therefore be important to avoid large interviewer workloads (to avoid such clusters being too large), although this has to be balanced against quality, consistency and cost issues of having too many interviewers working on a particular survey. In practice, the capacity of a single interviewer, who covers a specific cluster, tends to limit the practical size of a cluster, typically to what an interviewer can do in the few days over which he/she works on the particular project in hand.

COST-BENEFIT ANALYSIS

Harris (1977) illustrates that the cost, C , of conducting fieldwork in a two-stage design is given by the formula: (See [Equation 7](#))

where n is the number of primary sampling units (i.e. clusters); c_1 is the costs associated with each primary sampling unit (e.g. recruiting, training and supervising interviewers, sampling); nm is the total sample size; and c_2 represents the costs associated with each single interview (sampling from electoral register, interviewers' fees, travel between sites, etc.).

At one extreme, if all of the fieldwork were conducted at a single site/cluster, then the cost function would reduce to $(c_1 + mc_2)$. At the other extreme, the cost of running a completely unclustered SRS would be: (See [Equation 8](#))

as there would be as many primary sampling units as the overall sample size.

What this is indicating is that apparently saving money in travelling and area set-up costs by using a clustered sampling design with a small number of clusters might be a false economy. [Figure 5](#) shows that, although the actual cost of each interview increases steadily with the number of clusters (n), the effective cost decreases. A marginal cost of each interview (c_2) of £8 and a cluster set-up cost (c_1) of £800 has been assumed. The actual cost of using 40 clusters and 250 respondents per cluster comes out at £11.20 per actual interview. However, to achieve the same level of precision as a single (unclustered) interview would cost about five times as much (i.e. £57). *The definition of the effective cost per interview is the actual cost multiplied by the DE* (and the DE in the forty-cluster situation in the example described earlier is 5.1).

Effective costs can therefore be very great where the respondents are highly clustered. A clustered survey (with a DE of 2.0) of 1000 interviews costing £12,000 in terms of fieldwork would yield the same amount of precision as an unclustered survey of 500. Although the cost per actual interview would still be £12, the cost per effective interview would be £24 (i.e. £12,000 divided by the effective sample of 500).

OTHER PRACTICAL EXAMPLES

Voting Intention (VI) for a Minority Party

A very small percentage of people vote for the party in question (0.56%), so to measure this accurately would take a very large sample. This involved putting together two years' of face-to-face fortnightly-wave Omnibus data to achieve a sample of 96,400. Fieldwork was conducted over 230 geographic Omnibus units nationally (of which there are 4900 in the total population), and an average of 415 respondents per unit. Had this been a simple random sample of 96,400, the VI would be precise to within 0.05% points. However, the DE for this method was calculated as 4.50 and so the effective size was 21,400 and the result precise to 0.1%. Here, the effective sample size is roughly a quarter of the actual sample size, and hence the CI is roughly twice (i.e. the square root of 4) as wide.

Using the simulation techniques described above, the effects of using larger geographical areas as clusters (i.e. sampling across 230 constituencies) and having a larger number of ward-based clusters were investigated, and the results are summarised in [Table 3](#).

Scenario 1 is what has just been described and involves using 230 ward-based clusters. With scenario 2, it is found that allowing the samples to cover a broader geography improves the DE and hence the effective sample size and the CI, although not dramatically so. This may be because a higher proportion of the population is potentially within the sampling frame and not excluded as non-sampled clusters (e.g. 230 of 641 constituencies = 35% of the population, as compared to 230 of approximately 10,000 wards = 2.3% of population). The DE may be improved further, in practice, because with the clusters of constituencies being so much larger than wards, it is likely that average constituency survey measures are more similar to each other than average ward measures (ρ is likely to be smaller *between* constituencies than wards, hence the DE is smaller). Finally, scenario 3, which involves using the same types of cluster units (i.e. wards) as scenario 1 but twice as many of them, yields a considerably improved DE and effective sample size. The effect of increasing the numbers of clusters has been described in detail in an earlier section. Tipping and Pickering (2004) carried out some work to look at how the precision of survey estimates, design effect and ρ values for health measures changes with the size of the cluster. It is possible that ρ values can potentially double as the cluster size is halved.

Local Authority Satisfaction Survey

A recent large-scale, ad hoc survey of 14,000 respondents across 35 LAs was conducted involving random sampling of 400 people within each selected LA. With this design, six interviews per day could be averaged, thus giving a total fieldwork cost of £315,000 or £22.50 per interview. The DE for this was 5.9, giving an effective cost of £133 per interview. Covering twice as many LAs (70) reduces the average number of daily interviews that would be achieved (to 5.25) and so the cost of interviewing 14,000 respondents increases to £367,000 (or £26.20 per interview). However, for this design, the DE is only 2.8, so the effective interview cost comes out at £73. Therefore, although it apparently looks like the 35-cluster situation is costing 14% less than the 70-cluster design, the cost *per effective interview* is actually almost double.

CONCLUSIONS

The first point of interest to researchers may be that there is a price to pay for the convenience of running a clustered sample in order to cut down on time and costs in travelling between interviews. This price comes in the form of a reduced amount of statistical reliability in the sample, a widening of the CIs and a reduction in the effective sample size as compared to an equivalent simple random sample. The effect of clustering can be considerable.

A mathematical formula is available to enable the CI and DE for clustered samples for proportion measures to be calculated. At first sight, this formula may seem quite complex, needing a large amount of statistical information about the sample design in order to calculate it. Many researchers use this as an argument for ignoring the issue. All one would need is the sampling point/cluster/interviewer identifier attached to each record: no new data at all. The main point to be aware of is that there are two components of variance in a clustered sample which add together: the *within* and the *between* components.

The DE (which can vary from question to question within a survey) for clustered samples is heavily dependent on the amount of cluster-to-cluster differences in the results. Unfortunately, this is often not known until after the survey has been carried out and the result obtained, thus making accurate predictions of statistical reliability in proposals very difficult to make. This is compounded by the fact that this is likely to differ substantially from one question in the survey to another. To help get around this, it may be of value to look at the results of past surveys which asked similar types of questions. If very little historical data are forthcoming, rather than quoting (at

proposal stage) the survey's statistical reliability as if it were a simple random sample, it would be wiser for researchers to imply that the SRS estimate is almost never valid. At the bottom line, one could potentially consider building-in a p of up to 0.2, even if only due to interviewer effects. If one is concerned that someone else might offer more precision through ignoring the facts, at least one could flag up the general issues around clustering so that the research purchaser becomes aware of this and the agency demonstrates that it has thought about these issues where others may not have.

The DE is also highly dependent on the number of clusters covered in the survey, thus making it virtually impossible to compensate a low-cluster coverage survey by having a large sample size *within* a cluster. This highlights the importance of covering as many different geographic areas across the target area as one can afford to, even if it means only sampling very few individuals *within* each.

Although there is still a valid (even inevitable) place for cluster sampling for practical and cost reasons, researchers should be aware that a balance must be struck. In such circumstances, it may be worth considering an increase in the number of clusters sampled such as sampling from, say, 40 clusters with 20 respondents in each, rather than 20 clusters with 40 in each. Doing this can have a considerable effect on increasing precision, increasing the effective sample size, and indeed on the real cost per interview.

Admittedly, much of material presented here is not new, but issues remain, perhaps even more so these days, as companies and agencies aim to be commercially viable and produce results sooner and more cheaply than their competitors to increase profit margins. A key aim of this paper is to at least revive discussion of this topic among providers and users of survey research. This also has an implication for deciding which mode of interview to use as telephone and postal surveys would not generally suffer from the type of clustering described here. The issue of clustering can be considered as one of the areas where researchers' quests for quicker and cheaper results can erode the high levels of statistical input and technical standards adopted.

APPENDIX: THE SPREAD OF RESULTS *WITHIN* EACH CLUSTER (SATISFACTION SURVEY) (See [Table 4](#))

REFERENCES

- Cochran, W.G. (1968) *Sampling Techniques*, 2nd edn, pp. 66–68 and pp. 246–248. New York: Wiley.
- Collins, M. & Butcher, B. (1983) 'Interviewer and clustering effects in an attitude survey', *Journal of the Market Research Society*, **25**, pp. 39–58.
- Collins, M. & Goodhardt, G. (1978) Value for money in research design, paper presented at the Proceedings of the MRS Annual Conference, Brighton.
- Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953) *Sample Survey Methods and Theory*. New York: Wiley.
- Harris, P. (1977) 'Effect of clustering on costs and sampling errors of random samples', *Journal of the Market Research Society*, **19**, 3, pp. 112–122.
- Journal of the Market Research Society*, **25**, 1.
- Kish, L. (1965) *Survey Sampling*. New York: Wiley.
- Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T. (2000) *A Users Guide to ML-wiN*. University of London.
- Tipping, S. & Pickering, P. (2004) Impact of the geographical size of clusters on the precision of survey estimates. *Survey Methods Newsletter*, **23**, Winter.

NOTES & EXHIBITS

$$\bar{p}$$

EQUATION 1

$$CI = \pm 1.96 * \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

EQUATION 2

$$DE = \frac{n_{actual}}{n_{effective}}$$

EQUATION 3

$$DE = \left(\frac{SE_{cl}}{SE_{SRS}} \right)^2 = \frac{SE_{cl}^2}{\left(\frac{\bar{p}(1 - \bar{p})}{n} \right)} \quad (3)$$

EQUATION 4

$$v(\bar{p}) = \frac{1 - f_1}{n(n - 1)} \sum_{i=1}^n (p_i - \bar{p})^2 + \frac{f_1(1 - f_2)}{n^2(m - 1)} \sum_{i=1}^n p_i q_i$$

$$[\sum_{i=1}^n (p_i - \bar{p})^2]$$

$$[[\sum_{i=1}^n (p_i q_i)]$$

EQUATION 5

$$CI = \pm 1.96 * \sqrt{\frac{p(1-p)}{n}}$$

FIGURE 1: CALCULATION OF CI AND DE WITH CLUSTERING

No. clusters in sample (n)	30	
No. clusters in pop (N)	450	
GB population 16+	45,700,000	
Population per cluster (M)	101,556	
Sample size	10,000	
Mean cluster size (m)	333	
p (estimate)	79.47%	
S. frac – clusters (f1)	0.067	
S. frac – within clusters (f2)	0.003	
$\sum (p_i - p)^2$	0.105	
$\sum (p_i q_i)$	4.627	
"Between" component	0.0001125	
"Within" component	0.0000010	
Variance of p (= within + between)	0.0001136	
CIHW: +/-	2.09%	
DE	6.960	
Effective n	1437	

Assuming SRS nationally	
p (est)	79.5%
Sample size	10,000
Variance of p	0.0000163
CIHW: +/-	0.79%

Ratio = DE

$$v(\bar{p}) = \frac{1-f_1}{n(n-1)} \sum_{i=1}^n (p_i - \bar{p})^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_{i=1}^n p_i q_i$$

EQUATION 6

$$DE = 1 + (m - 1)\rho$$

TABLE 1: DESIGN EFFECT CHANGES WITH CLUSTER SIZE AND INTRA-CLUSTER CORRELATION

Design effects	Intra-cluster correlation (ρ)					
	0.01	0.02	0.03	0.05	0.10	0.2
Average cluster size (m)						
10	1.09	1.18	1.27	1.45	1.90	2.8
20	1.19	1.38	1.57	1.95	2.90	4.8
30	1.29	1.58	1.87	2.45	3.90	6.8
40	1.39	1.78	2.17	2.95	4.90	8.8
50	1.49	1.98	2.47	3.45	5.90	10.8
100	1.99	2.98	3.97	5.95	10.90	20.8
200	2.99	4.98	6.97	10.95	20.90	40.8

Source: Collins and Goodhardt (1978)

TABLE 2: EFFECTIVE SAMPLE SIZE CHANGES WITH CLUSTER SIZE AND INTRA-CLUSTER CORRELATION

Effective sample size/cluster	Intra-cluster correlation (ρ)					
	0.01	0.02	0.03	0.05	0.10	0.20
Avg cluster size (m)						
10	9	8	8	7	5	4
20	17	14	13	10	7	4
30	23	19	16	12	8	4
40	29	22	18	14	8	5
50	34	25	20	14	8	5
100	50	34	25	17	9	5
200	67	40	29	18	10	5
Maximum	100	50	33	20	10	5

Source: Collins and Goodhardt (1978)

FIGURE 2: DE CHANGES WITH CLUSTER COVERAGE

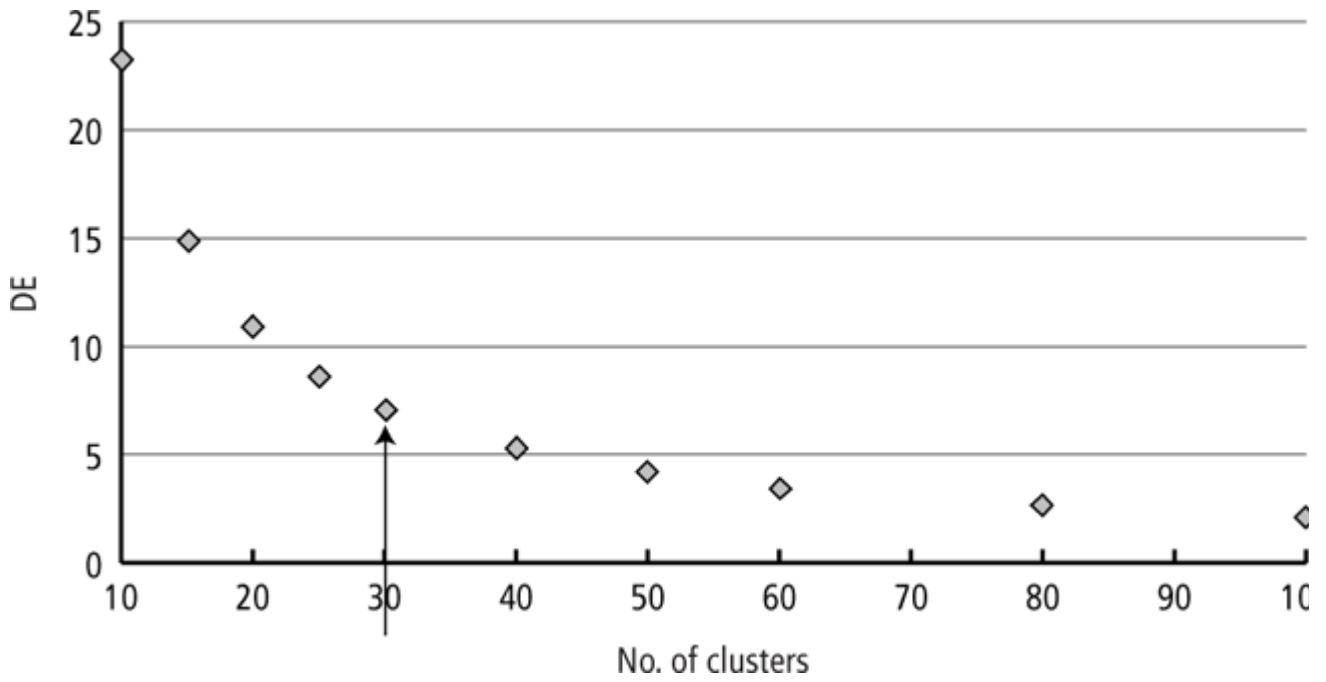
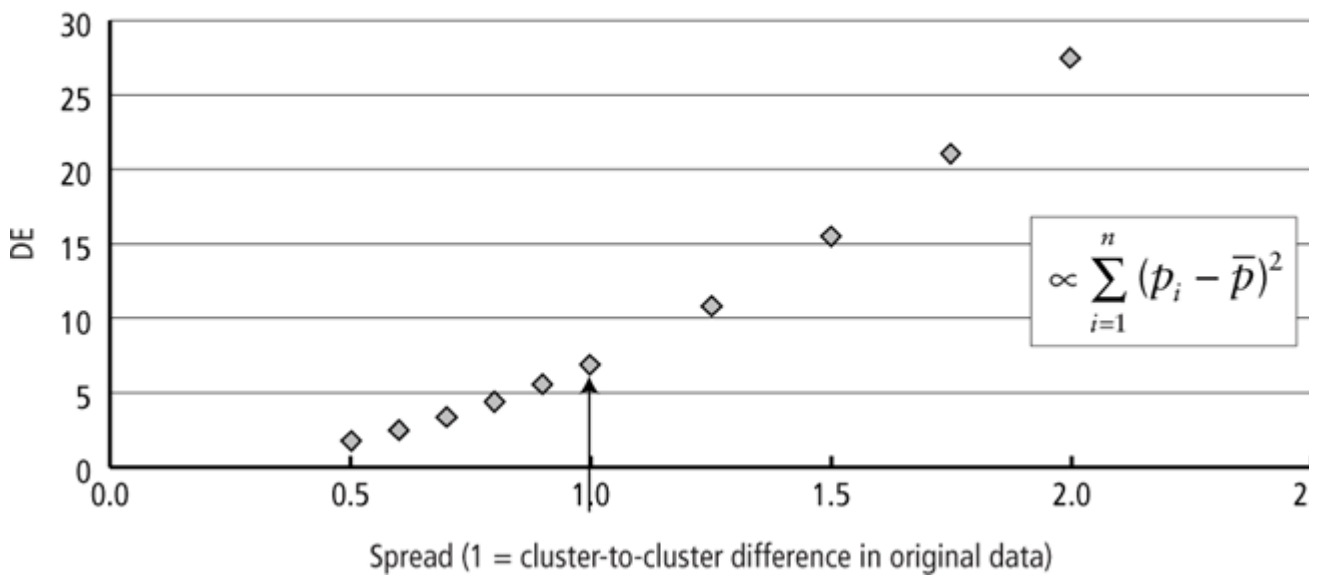


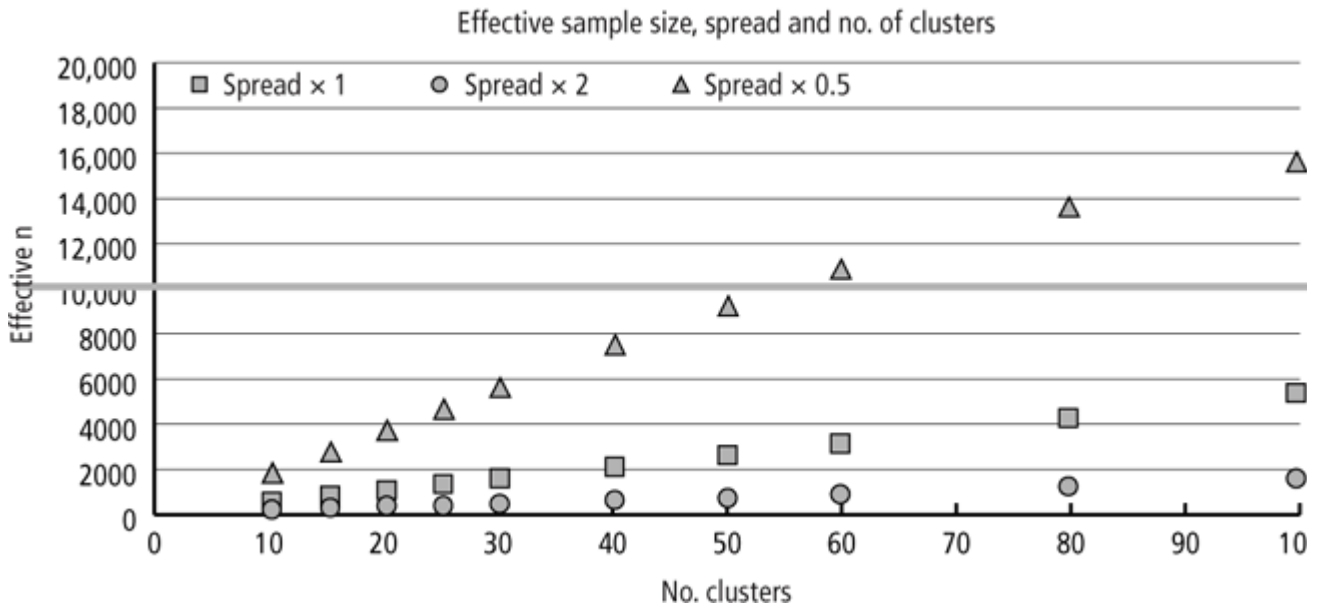
FIGURE 3: DESIGN EFFECT CHANGES WITH CLUSTER DIFFERENCES

DE with between-cluster spread of results – 30LAs * 333 respondents each



$$\sum_i (p_i - \bar{p})^2$$

FIGURE 4: DESIGN EFFECT CHANGES WITH CLUSTER NUMBERS AND DIFFERENCES



EQUATION 7

$$C_{SRS} = nmc_1 + nmc_2 = nm(c_1 + c_2)$$

EQUATION 8

$$C = nc_1 + nmc_2$$

FIGURE 5: ACTUAL AND EFFECTIVE INTERVIEW COSTS WITH NUMBER OF CLUSTERS

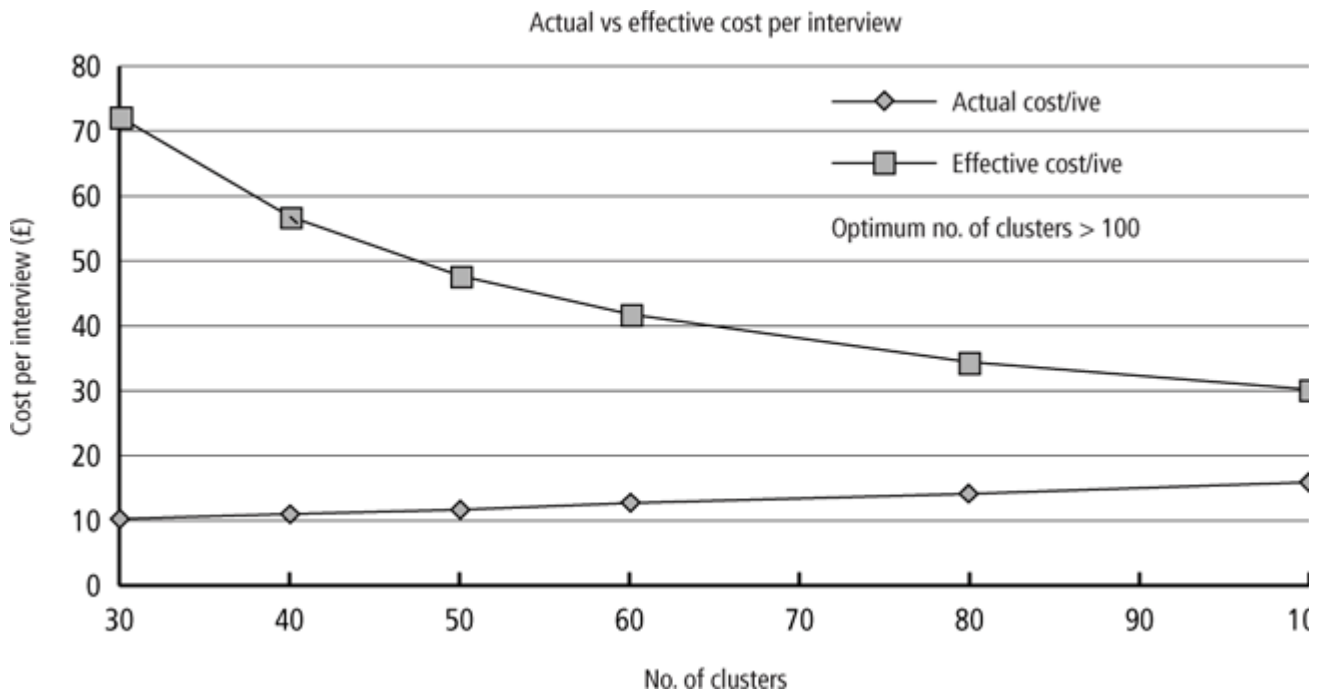


TABLE 3: EFFECT OF CHANGING THE GEOGRAPHY OF CLUSTERING

	DE	CI half width (%)	Effective i
Scenario 1: 230 ward clusters	4.58	0.102	21 024
Scenario 2: 230 constituencies	3.36	0.087	28 690
Scenario 3: 460 ward clusters	1.13	0.051	85 030
Simple random sample	1.00	0.048	96 389

TABLE 4

Unit	Original data (%)	Spread factor = 2 (%)
Cluster 1	85.1	90.7
Cluster 2	77.9	76.4
Cluster 3	68.0	56.4
Cluster 4	84.3	89.2
Cluster 5	83.2	86.9
Cluster 6	84.2	88.9
Cluster 7	83.9	88.4
Cluster 8	75.2	71.0
Cluster 9	80.7	81.9
Cluster 10	83.7	87.9
Cluster 11	68.2	56.9
Cluster 12	82.4	85.3
Cluster 13	83.4	87.3
Cluster 14	84.5	89.5
Cluster 15	74.8	70.1
Cluster 16	79.7	80.0
Cluster 17	80.5	81.5
Cluster 18	87.2	94.8
Cluster 19	87.3	95.1
Cluster 20	84.6	89.7
Cluster 21	78.0	76.6
Cluster 22	85.0	90.5
Cluster 23	86.4	93.3
Cluster 24	78.9	78.3
Cluster 25	62.9	46.3
Cluster 26	77.1	74.7
Cluster 27	79.5	79.6
Cluster 28	85.6	91.8
Cluster 29	83.0	86.5
Cluster 30	76.6	73.8
Overall	79.5	79.5

© Copyright World Advertising Research Center 2005
World Advertising Research Center Ltd.
Farm Road, Henley-on-Thames, Oxon, United Kingdom, RG9 1EJ
Tel: +44 (0)1491 411000, Fax: +44 (0)1491 418600

All rights reserved including database rights. This electronic file is for the personal use of authorised users based at the subscribing company's office location. It may not be reproduced, posted on intranets, extranets or the internet, e-mailed, archived or shared electronically either within the purchaser's organisation or externally without express written permission from World Advertising Research Center.



www.warc.com