# The Impact of Cluster (Segment) Size on Effective Sample Size

Steven Pedlow, Yongyi Wang, and Colm O'Muircheartaigh
National Opinion Research Center, University of Chicago

## Abstract

National in-person (face-to-face) interview surveys usually use a sample design called area probability sampling. Clusters of interviews located close to each other tend to be correlated, as measured by the intraclass correlation, often represented by the Greek letter rho (ρ). Design decisions have an impact on the intraclass correlation, and the intraclass correlation has an impact on the variability of sample estimates. We used data from the Making Connections study, sponsored by the Annie E. Casey Foundation, to study the effect of cluster sizes (the total number of housing units in the cluster, and not the number of interviews per cluster) on sampling variances and effective sample sizes. It is logical that increasing the cluster size would decrease clustering effects because people in larger groups are less similar than people in smaller groups, and this paper tries to measure this effect. Section 1 provides background on area probability sampling. Section 2 provides the theoretical groundwork for this paper on cluster size effects. The NORC National Frame, which inspired this work, is described in Section 3. Section 4 introduces the Making Connections data set we analyzed. Sections 5-7 finish by describing our methods, results, and conclusions, respectively.

**Keywords**: Rho, Intraclass correlation, area probability sampling, Making Connections.

## 1. Area Probability Sample Designs

Planning for a national sample of housing units for in-person (face-to-face) interviews is a balancing act of cost versus variance. Statistical (variance) optimality is achieved by a simple random sample, but it is impractical (cost) to send field interviewers across the whole nation. Therefore, most national samples involving in-person interviewing use area probability sampling. This sampling strategy involves selecting clusters across the country, then selecting a subsample of housing units from each selected cluster. Clustering in this way saves costs, but increases the variance of sample estimates because people who live near each other tend to be more similar to each other than they are to the rest of the country.

Many area probability samples have three stages of sampling. The first stage of sampling selects county-level units (counties are often combined, especially within metropolitan areas), sometimes referred to as primary sampling units (PSUs). The second stage selects neighborhoods or blocks within the selected primary sampling units. These selections are sometimes referred to as secondary sampling units (SSUs). The first two stages are often selected using Census data, which gives the number of housing units down to the block level. The third stage, however, often involves listing all the housing units in the selected SSUs so that housing units can be selected.

Many variations are possible as many different decisions could be made in terms of how many PSUs, SSUs, and housing units should be selected. There are also decisions to be made in how large the units at each stage are made (e.g., how to combine the counties). The focus of this paper is on how large to make the SSUs.

## 2. Cluster Size Effects

Secondary sampling units (SSUs) are sometimes referred to as segments, but this paper will refer to them as clusters. In this paper, we ignore the aspects of the sample design other than the size of the SSUs. We assume that there is a given number of interviews per cluster, which we will represent by b. It is very important to separate this from the cluster size, which is the population for the last stage of selecting housing units. Then, b is the sample size for this stage of sampling. So, given the sample size within cluster, what effect does the cluster size have?

The variability of sample estimates depends on b and the intraclass correlation, which measures how similar people within the same cluster are compared to how similar they are to the people in other clusters. In fact, there is a well known approximation to the relationship of the effective sample size given here:

$$n(eff\ ) \approx \frac{n}{\left[1 + \rho(\bar{b} - 1)\right]},$$

where $\overline{b} = n/(\#SSUs)$ is the average number of interviews per cluster. Note that if ρ is 0 or $\overline{b}$ = 1 (simple random sampling has no clustering), the effective sample size is the same as the total sample size . The intraclass correlation (ρ) is generally positive, so the effective sample size is generally less than the total sample size.

The intraclass correlation differs for each variable, depending on how similar people who live near each other are on any particular characteristic. Financial characteristics tend to have high intraclass correlations because people who live near each other tend to have similar financial situations. Behavioral variables, however, tend to have smaller intraclass correlations because even people in similar financial situations tend to have different opinions and behaviors. Larger intraclass correlations reduce the effective sample size and the key objective of this paper is to measure the extent that the intraclass correlation can depend on the cluster size, which will be represented by $\overline{N}$ . Of course, not every cluster will have the same size, so the average cluster size ( $\overline{N}$ ) is used.

Hansen, Hurwitz, and Madow (1953) suggested a logarithmic relationship between the average cluster size and the intraclass correlation:

$$\rho = a\left(\overline{N}\right)^m,$$

where a and m are different parameters for different variables. The assumption in the formula is that m < 0 so that as the average cluster size increases, the intraclass correlation will decrease. This is explained by the fact that smaller areas tend to be more similar than larger areas. As a very simple example, one neighborhood in Miami will tend to be more similar than the entire city, which will tend to have diversity when all the different neighborhoods are combined.

Getting back to our question, how large should clusters be? Larger clusters will have smaller intraclass correlations and therefore, higher effective sample sizes. However, larger clusters increase costs because of travel within the cluster and the simple technique of having to list more housing units in larger clusters.

## 3. NORC National Frame

NORC's national frame, revised after every Decennial Census for use in that decade, is an area probability sample with three stages. We have gradually increased the minimum cluster sizes over time. The 1990 NORC National Frame used a minimum cluster size of 50, while other NORC studies in the 1990's often used a minimum of 100. The 2000 NORC National Frame, however, has made use of available city-style postal listing files. Using these lists, NORC now uses entire census tracts as clusters, which average about 2000 housing units. Traditional listing is still used, but only in rural areas where city-style addresses are not available. For these rural areas, a minimum cluster size of 300 was used.

## 4. Making Connections Data

To study different cluster sizes, we used data from the Making Connections project. Making Connections is a study funded by the Annie E. Casey Foundation to study children and families in deprived communities within 10 United States cities, which are listed here:

- Denver, Colorado
- Des Moines, Iowa
- Indianapolis, Indiana
- San Antonio, Texas
- Seattle, Washington
- Hartford, Connecticut
- Louisville, Kentucky
- Milwaukee, Wisconsin
- Oakland, California
- Providence, Rhode Island

Since these were inner city areas, we were able to obtain the frame of city-style addresses for the entire study areas. It is important to note that Making Connections sample selection did not use any clustering; simple random sampling was used within each community. This is what allows us to use the data to study different cluster sizes. Since the interviews are a simple random sample, we can form clusters of "any" size and calculate the intraclass correlation for the clusters formed using the interview data collected. If clustering was used for sampling, we would have many empty clusters depending on how we formed them, and this work would not have been possible.

It is important to note that there is sampling error involved in estimating the intraclass correlations just as for any other sample parameter. In particular, larger clusters will typically have more interviews and therefore, better intraclass correlation estimates. We analyzed 18 important Making Connections variables, but this paper presents data from only one community: Louisville, Kentucky. Here is a list of the 18 variables we studied (* indicates a binary variable):

- Income (11)
    - Wages/Salary
    - Commissions? *
    - Self-Employment Income? *
    - Interest Income? *
    - Social Security Income? *
    - Supplemental Security Income? *
    - Public Assistance? *
    - Retirement Pensions? *
    - Other Work Income? *
    - Veterans Income? *
    - TOTAL Household
- Months Lived in Louisville
- Volunteer in Last 12 Months? *
- Postpone Prescription (RX) in Last 12 Months? *
- Missed Mortgage Payment in Last 12 Months? *
- Housing (3)
    - Own your own house? *
    - Total Rent
    - Total Payment

Most of the 18 variables are financial, and 13 of them are binary variables.

## 5. Cluster Size Choices

Altogether, Louisville had 704 interviews among 7236 housing units. The context for this paper is the set of decisions that we make in designing a sample. Were we to select a two-stage sample from Louisville, we would need to define in advance the primary sampling units. We assume for this paper that a selected primary sampling unit is comprised of the Louisville neighborhoods from the Making Connection project. The next step would be to subdivide the primary sampling unit into clusters. In creating these clusters, we describe 47 scenarios based on combinations of two parameters. First, an "alone" (set-aside) parameter specified blocks that were large enough (according to the 2000 Decennial Census) to be clusters on their own. Typical values for "alone" were: no set-asides, 50, 100, 150, and 200. From the remaining blocks, we put consecutive blocks together until they formed a cluster of at least the minimum size, which ranged from 25 to 2000 (11 different sizes). Typically, four or five different set-aside values were used for each minimum size. The contribution of the "alone" parameter is that it can prevent long strings of small blocks from being combined with a block almost as large as the minimum size (e.g., 20, 20, 20, 20, 20, 20, 20, and 150 when the minimum is 200).

## 6. Results

For each of the 18 variables, we estimated the intraclass correlation for the clusters under each of the 47 scenarios outlined above. Figures 1 and 2 show scatter plots of intraclass correlation size and average cluster size for two variables. The average cluster size, of course, is simply 704 divided by the number of clusters formed. This does not separate the effects of the minimum and set-aside factors, but is a helpful simplification. Figure 1 shows the scatter plot for the binary variable "Missed Mortgage Payment in Last 12 Months." The pattern seems to be clearly downward with a curve, which actually fits the logarithmic relationship given by Hansen, Hurwitz, and Madow quite well. Figure 2 shows a less clear picture for "Income from Wages and Salary," but the intraclass correlations still tend to be smaller for the larger average cluster sizes.

Table 1 below shows the mean intraclass correlation over all 47 scenarios for all 18 of the variables. This table also includes the number of observations for the variables (some of which are asked only of a subset of the 704 interviews):

Table 1. Mean intraclass correlations

| Variable | n | Mean |
|---|---|---|
| Total Rent on Unit | 498 | 0.0259 |
| Amount of Mortgage/Rent Paid | 572 | 0.0168 |
| Own your own house? | 693 | 0.0159 |
| Total Household Income | 634 | 0.0154 |
| Income From Wages or Salary | 701 | 0.0106 |
| Income From Social Security | 696 | 0.0064 |
| Income From Supplemental Security | 689 | 0.0046 |
| Months Lived in Louisville | 694 | 0.0038 |
| Income From Commissions | 695 | 0.0029 |
| Income From Retirement | 696 | 0.0028 |
| Missed Mortgage Payment in Last 12 Months | 703 | 0.0024 |
| Income From Other Work | 698 | 0.0019 |
| Income From Public Assistance | 697 | 0.0018 |
| Income From Self-Employment | 693 | 0.0017 |
| Income From Veterans Pay | 697 | 0.0015 |
| Income From Interest | 693 | 0.0011 |
| Volunteered in Last 12 Months | 702 | 0.0009 |
| Postpone RX in Last 12 Months | 701 | 0.0005 |

These intraclass correlations are actually quite a lot smaller than our national surveys. The reason is straightforward – the population is only one inner-city area. Because of this, the similarity is not much stronger within clusters than within the whole population. This results in low intraclass correlations, which has consequences for our study.

If we recall the logarithmic relationship from Hansen, Hurwitz, and Madow, we can take the natural logarithm of each side to produce:

$$\ln(\rho) = \ln(a) + m * \ln(\overline{N}) \quad .$$

Linear regression is now possible, with m as the slope of the regression line. Of course, on our scatter plots (see Figures 3-4), these regression lines are curves.

Figure 3 shows one example for the variable, "Do You Own Your Own House?" The line does seem to follow the pattern of the data. Figure 4 shows a counter-intuitive regression line for the "Months Lived in Louisville" item. For 2 of the 18 variables, we actually found that the intraclass correlation significantly increased with the average cluster size. We hypothesize that this is simply due to random variation in the measurements of the intraclass correlation, and is bound to happen for some percentage of variables examined. We also believe that this is more likely in cases such as this where the intraclass correlations are localized and small.

Table 2 contains the regression parameters for all 18 variables, as well as intraclass correlation regression estimates for three different average cluster sizes. The average cluster sizes in Table 2 reflect the minimum cluster sizes used in recent NORC history. Minimums of 50 and 300 resulted in average cluster sizes of 83 or 381, while $\overline{N}$ = 2000 refers to our new method of using entire census tracts as clusters. As shown in Figure 4, "Months in Louisville" was one of the two variables that had a significantly positive slope; the other was "Income from Retirement." A third variable, "Volunteered in the Last 12 Months" had a non-significant positive slope. Of the other 15 variables, 3 had a non-significant negative slope, so a summary would be 12 significantly decreasing slopes, 2 significantly increasing slopes, and 4 non-significant slopes.

Table 3 shows effective sample sizes based on the estimated intraclass correlations given by our models. Since effective sample size also depends on the average number of interviews per cluster ( $\overline{b}$ ), we assumed $\overline{b}$ to be 5, which is typical for NORC studies. The effective sample sizes are also based on an idealized sample size of 700 rather than the actual number of observations for each variable. The last column in Table 3 shows the gain in effective sample size by increasing the average cluster size from 83 to 2000. It's apparent that the gains shown here are small. However, one key explanation is that the intraclass correlations during our work are so small, and are much smaller than national studies. These small intraclass correlations result in small losses, so there is not much room for improvement. For example, for "Income from Commissions," even with $\overline{N}$ = 83, the loss is only 10 from 700 to 690. By increasing $\overline{N}$ to 2000, 70% of these 10 are regained. Overall, the gain in effective sample size is only 1.03%.

Table 4 shows the losses due to clustering, and how much is gained back by increasing the average cluster size from 83 to 2000. Again, these numbers use an idealized total sample size of 700 and an average number of interviews per cluster equal to 5. Looking at "Income from Commissions," 72.6% of the loss is gained back by increasing the average cluster size. Looking at all 18 variables, the gains are all over the place, from 0% to almost 100%. Three losses are repressed because we can't see any possibility that the intraclass correlation would increase with average cluster size. We believe these gains could therefore be quite large in more typical national surveys with larger intraclass correlations. The larger the intraclass correlations, the more that can be gained by increasing the average cluster size.

## 7. Conclusions and Future Work

In conclusion, we had smaller intraclass correlations than most surveys. Because of this, we showed only small gains by increasing the average cluster size. However, by examining the loss due to clustering and how much could be gained back, we showed a significant gain. We intend to next use all 10 Making Connections communities in an effort to see if these gains hold up.

The increasing intraclass correlations are strange, and show the volatility in our intraclass correlation calculations. This might be due to the fixed sample size which allowed different precisions in our estimates of the intraclass correlations. However, we also feel that this kind of result is bound to happen if enough variables are studied, and it occurred for only 2 out of the 18 variables we studied.

## Reference

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. New York: Wiley.

Figure 1. Example scatter plot showing the relationship between intraclass correlation and average cluster size.

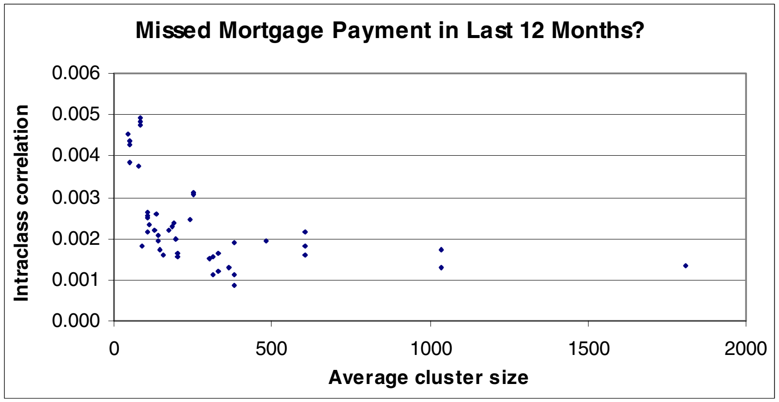

Figure 2. A second example scatter plot.

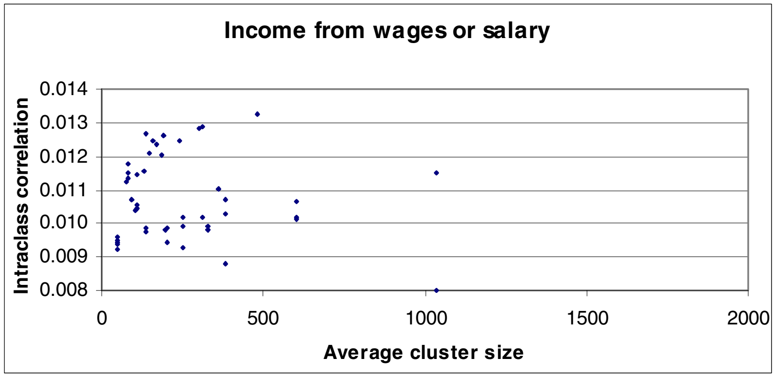

Figure 3. Example scatter plot with regression line superimposed.

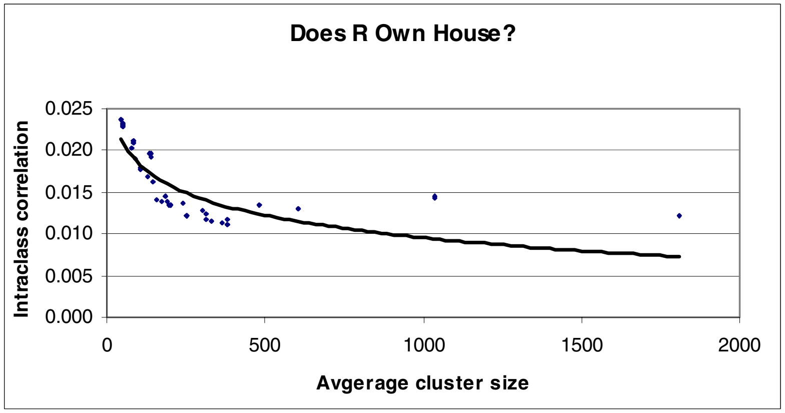

Figure 4. A scatter plot with a counter-intuitive regression line.

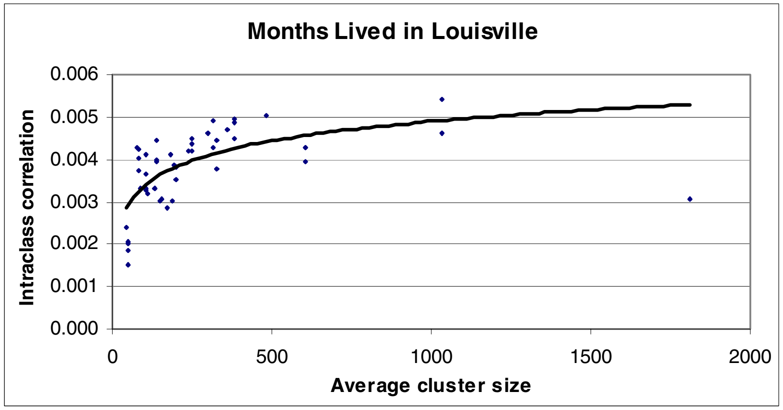

Table 2. Regression Parameters and Example ρ estimates for Several Average Cluster Sizes.

| Variable | Regression Parameters | | Regression Estimate for ρ | | |
|---|---|---|---|---|---|
| | a | m | $\overline{N} = 83$ | $\overline{N} = 381$ | $\overline{N} = 2000$ |
| Total Rent on Unit | 0.0600 | -0.1620 | 0.0293 | 0.0229 | 0.0175 |
| Amount of Mortgage/Rent Paid | 0.1044 | -0.3570 | 0.0215 | 0.0125 | 0.0069 |
| Own your own house? | 0.0523 | -0.2320 | 0.0188 | 0.0132 | 0.0090 |
| Total Household Income | 0.0356 | -0.1620 | 0.0174 | 0.0136 | 0.0104 |
| Income From Wages or Salary | 0.0166 | -0.0900 | 0.0112 | 0.0097 | 0.0084 |
| Income From Social Security | 0.0100 | -0.0890 | 0.0067 | 0.0059 | 0.0051 |
| Income From Supplemental Security | 0.0059 | **-0.0570** | 0.0046 | 0.0042 | 0.0039 |
| Months Lived in Louisville | 0.0013 | 0.2030 | 0.0031 | 0.0042 | 0.0059 |
| Income From Commissions | 0.0217 | -0.4100 | 0.0035 | 0.0019 | 0.0010 |
| Income From Retirement | 0.0008 | 0.2400 | 0.0022 | 0.0031 | 0.0047 |
| Mortgage Payment in Last 12 Months | 0.0155 | -0.3710 | 0.0030 | 0.0017 | 0.0009 |
| Income From Other Work | 0.0107 | -0.3700 | 0.0021 | 0.0012 | 0.0006 |
| Income From Public Assistance | 0.0015 | **-0.0070** | 0.0015 | 0.0015 | 0.0015 |
| Income From Self-Employment | 0.0071 | **-0.2850** | 0.0020 | 0.0013 | 0.0008 |
| Income From Veterans Pay | 0.7543 | -1.2510 | 0.0030 | 0.0004 | 0.0001 |
| Income From Interest | 0.0949 | -0.8790 | 0.0020 | 0.0005 | 0.0001 |
| Volunteered in Last 12 Months | 0.0002 | **0.2210** | 0.0006 | 0.0009 | 0.0013 |
| Postpone RX in Last 12 Months | 0.0051 | -0.4240 | 0.0008 | 0.0004 | 0.0002 |

Slope parameters in **bold** are non-significant at p=.05.

Table 3. Effective Sample Sizes Based on Regression Models

| Variable | Effective Sample Size* ($\overline{b} = 5$) | | | Gain<br>83 -> 2000 |
|---|---|---|---|---|
| | $\overline{N} = 83$ | $\overline{N} = 381$ | $\overline{N} = 2000$ | |
| Total Rent on Unit | 627 | 641 | 654 | 4.41% |
| Amount of Mortgage/Rent Paid | 644 | 667 | 681 | 5.69% |
| Own your own house? | 651 | 665 | 676 | 3.78% |
| Total Household Income | 654 | 664 | 672 | 2.69% |
| Income From Wages or Salary | 670 | 674 | 677 | 1.08% |
| Income From Social Security | 682 | 684 | 686 | 0.65% |
| Income From Supplemental Security | 691 | 688 | 684 | -1.10% |
| Months Lived in Louisville | 687 | 688 | 689 | 0.30% |
| Income From Commissions | 690 | 695 | 697 | 1.03% |
| Income From Retirement | 694 | 691 | 687 | -0.98% |
| Mortgage Payment in Last 12 Months | 692 | 695 | 697 | 0.83% |
| Income From Other Work | 694 | 697 | 698 | 0.57% |
| Income From Public Assistance | 696 | 696 | 696 | 0.01% |
| Income From Self-Employment | 694 | 696 | 698 | 0.48% |
| Income From Veterans Pay | 692 | 699 | 700 | 1.18% |
| Income From Interest | 695 | 699 | 700 | 0.73% |
| Volunteered in Last 12 Months | 698 | 698 | 696 | -0.25% |
| Postpone RX in Last 12 Months | 698 | 699 | 699 | 0.23% |

Table 4. Clustering Sample Size Reductions.

| Variable | Clustering Reduction in Eff(n) | | | Gain<br>83 -> 2000 |
|---|---|---|---|---|
| | $\overline{N} = 83$ | $\overline{N} = 381$ | $\overline{N} = 2000$ | |
| Total Rent on Unit | -73 | -59 | -46 | 37.6% |
| Amount of Mortgage/Rent Paid | -56 | -33 | -19 | 66.1% |
| Own your own house? | -49 | -35 | -24 | 50.4% |
| Total Household Income | -46 | -36 | -28 | 38.7% |
| Income From Wages or Salary | -30 | -26 | -23 | 24.1% |
| Income From Social Security | -18 | -16 | -14 | 24.2% |
| Income From Supplemental Security | -9 | -12 | -16 | * |
| Months Lived in Louisville | -13 | -12 | -11 | 16.3% |
| Income From Commissions | -10 | -5 | -3 | 72.6% |
| Income From Retirement | -6 | -9 | -13 | * |
| Mortgage Payment in Last 12 Months | -8 | -5 | -3 | 69.0% |
| Income From Other Work | -6 | -3 | -2 | 69.0% |
| Income From Public Assistance | -4 | -4 | -4 | 2.2% |
| Income From Self-Employment | -6 | -4 | -2 | 59.4% |
| Income From Veterans Pay | -8 | -1 | 0 | 98.1% |
| Income From Interest | -5 | -1 | 0 | 93.9% |
| Volunteered in Last 12 Months | -2 | -2 | -4 | * |
| Postpone RX in Last 12 Months | -2 | -1 | -1 | 74.0% |