Evaluation of the Effect on Cost and Variances of the Group Quarters Cluster Size

Daniel Sommers, Steven P. Hefter
US Census Bureau
Washington, DC 20233-8700

Abstract

The American Community Survey selects its large group quarters (GQ) sample by selecting clusters of ten persons from the GQ sampling frame. Cluster size influences both the variance of the estimates and survey costs. Reducing the number of people in a GQ sample cluster will improve the reliability of the estimates while simultaneously increasing data collection costs. In this paper we evaluate the variance reduction gains from decreasing the sample cluster size for large GQs and the expected impact on survey cost. Additionally, we calculate a cost/benefit factor in order to select the cluster size that offers the most benefit for the increased cost.

Keywords: American Community Survey, Group Quarters, sampling variance

1. Introduction

This research was conducted with the goal of evaluating the variance reduction gains from a reduced GQ cluster size, while simultaneously performing a cost-benefit analysis to find an optimal cluster size. GQs are facilities that are owned or managed by an entity or organization, where there is a group living arrangement for the people living or staying in that facility (U.S. Census Bureau, 2009). Correctional institutions, juvenile facilities, nursing homes and college dorms are examples of GQs, and the GQ population in any given GQ is the number of people living in the GQ. Current sampling methods assign clusters of expected person level interviews within these living quarters and thereby introduce an increase to the variance of the estimates over and above simple random sampling. Additionally, survey costs will increase with any decrease in the cluster size. This is because a decreased cluster size will increase the number of sample clusters, requiring Field Representatives (FRs) to make more trips to GQs. We wanted to find an optimal cluster size where cost increases give the most return in terms of a reduction in variance.

Two approaches were taken. Both approaches use a similar cost model but differ in their assumptions about how survey costs will increase with a decreasing cluster size. The assumptions used were meant to identify the extremes or end points in our optimization problem, and are dependent upon FR travel costs. The first approach assumed a separate FR trip for each sample cluster, while the second assumed a separate trip for every GQ in sample. For this second approach, it was understood that for the GQs receiving multiple sample clusters all the interviews would be conducted during one trip. Currently, a significant portion of the GQ sample is clustered by month, meaning that FRs visit these

GQ facilities during the same randomly assigned month. Also, the number of facilities that receive multiple FR visits which are spread throughout the year is relatively small. This aspect of the second approach is discussed in more detail later in the paper.

Since our cost model was dependent upon the number of FR trips, we were only interested in the projected number of FR trips in the cost analysis. Each cost model was examined against a simulated GQ sample. Both models were restricted to cluster sizes of one through ten, and conclusions were based on the current GQ survey cost structure.

An optimal cluster size was selected for each approach, based on the minimum cost-variance product. Since these were assumed to be the extremes, the final optimal cluster size was selected as the cluster size that fell between these two points.

2. Background

The ACS GQ sample is made up of people in large and small GQs. Large GQs have an expected population of greater than 15, and small GQs have an expected population of less than or equal to 15. Clusters are assigned to the large GQs with probability proportional to size, and sample clusters of ten expected interviews are selected (U.S. Census Bureau, 2009).

A larger cluster size with a fixed overall sample would result in a smaller first stage sample. By contrast, reducing the cluster size with a fixed overall sample would increase the first stage sample. Additionally, the lowest cluster size, one, would produce a simple random sample of the GQ population, in effect changing the sample design to a one stage sample (Cochran, 1977). This relationship led to an examination of ACS design effects, where we observed the ratio of the variance from our current clustered sample to the variance from a simple random sample. Large design effects indicate a relatively large variance when compared to estimates from a simple random sample. A design effect for this study was estimated using ACS data.

The second aspect of this study involved the similarity among GQ residents. Many of the ACS GQs tend to have populations with similar characteristics. It follows that residents of the same GQ are more likely to answer the survey in a similar way. If this tendency is very high, and the sample cluster size is large, then an increasing portion of the overall variance of an estimate will be accounted for by group membership.

By decreasing the ACS cluster size we attempt to lessen the impact of this issue for the survey. As mentioned, decreasing the cluster size will increase the number of first stage sample. This would result in selecting fewer members who belong to the same group while increasing the number of sample groups.

In 2005 an initial inquiry into the effects of sample cluster sizes, the intraclass correlation coefficient and their effect on variances was implemented. An optimization method was not yet developed. Since reliable cost data was not yet available, field costs were not taken into consideration until a 2007 study, along with a cost equation and an optimization approach using both small and large GQs with 2006 cost data. The study was then revisited in 2008 looking only at large GQs and results completed in 2009 using updated cost data for fiscal year 2009.

3. Methodology

The key components of the optimization problem involve an approximate variance of a sample estimate and a cost model that estimates changes in survey cost. For this study, both of these formulas have cluster size *M* as the independent variable. An assumption in the variance calculations is that the overall sample size will not change. This in turn reduces the change in the variance component to an expression of the design effect that is dependent upon the cluster size, and the intraclass correlation coefficient, ρ. The intraclass correlation coefficient was estimated using 2006 ACS data. Since more recent design effects were identified as being much larger than those from 2006, an analysis using a range of values for the intraclass correlation coefficient was conducted.

For the cost model, two models were explored. Both models express a projected total cost per sample person that is dependent upon *M*. The product of the design effect and each cost model was then minimized with respect to *M*. This was done for the two approaches (discussed earlier) and an optimal cluster size was then identified.

3.1 Approximate Variance of a Sample Estimate, the Design Effect and Estimating ρ

The approximate variance of a sample estimate for the current GQ sample design was calculated using the following equation.

$$\text{var}(\bar{y}) = \left[(1-f)\left(s^2/n\right)\right]\left[1+(M-1)\rho\right]$$

Here we expressed *n* as the total sample of GQ persons. For this study we assumed the total sample size and population would not change, therefore the design effect (DE), which is expressed as $[1+(M-1)\rho]$, had the most impact on the cost benefit analysis (in terms of influence from variance). Since we were interested in changes in *M* for a fixed sample size, we needed only to find an estimate ρ and then express the design effect as a function of only *M*. This was then used in the optimization analysis instead of the entire variance expression.

The intraclass correlation coefficient was estimated by assigning the maximum of five design factors for each state as that state's design factor. An average was then calculated for all states resulting in an estimated DE of 1.96. Using this we estimated $\rho = 0.1067$.

3.2 Method for Developing Survey Cost Model One

Several assumptions were made for the first approach (model one). First we assumed that each sample cluster would receive a separate FR visit. We also assumed the number of FRs, the sample size and other administrative costs would remain the same and used this information to construct a total current cost per sample person variable called *F*. By making this assumption about the number of FRs remaining constant, we must recognize a potential limitation. For this model, when the cluster size is reduced, and the number of trips increase, we would likely need to hire more FRs. This is particularly true for very small cluster sizes. For our analysis we assumed that the number of FRs would remain fixed.

Additionally, it is important to note that the total GQ costs for this study only included costs at the regional office level, and did not include administrative costs at headquarters, printing, postage and data capture.

F was calculated by taking the total GQ survey cost for the regional offices and dividing it by the total number of completed GQ cases for fiscal year 2009 (U.S. Census Bureau, 2009). This means that *F* also includes current travel costs for the survey. After defining the total current cost per sample person, we then assumed that any additional sample clusters for the survey would increase the total survey cost. This would result in travel costs having the most impact on our cost model, and thereby introduced a travel cost per cluster variable (*T*). If the cluster size were to be reduced from a size of ten, the resulting increase in sample clusters would necessitate more FR trips. We expressed this relationship as:

$$\text{new clusters} = \left(\frac{sample}{M} - \frac{sample}{10}\right) = \left(\frac{n}{M} - \frac{n}{10}\right)$$

where $0 < M \leq 10$.

We used these assumptions and this expression to construct a projected total survey cost (C_t) formula:

$$C_t = (\text{total current cost}) + (\text{new clusters}) \times (T)$$

We use total current cost for this expression and not *F* since this expression represents the projected cost for the entire survey. This expression can then be used to derive a projected cost per sample person as follows.

$$C_t = \left[(\text{total current cost/sample person}) + \left(\frac{(\text{new clusters}) \times (T)}{n}\right)\right] \times (n)$$

n is used since this represents the total number of people in sample. The total current cost per sample person is already defined as *F*, and new clusters has been expressed above. This allows the following.

$$C_t = \left[F + \frac{\left(\frac{n}{M} - \frac{n}{10}\right)T}{n}\right] \times n = \left[F + \left(\frac{1}{M} - \frac{1}{10}\right)T\right] \times n$$

The projected total survey cost per sample person (C_{pp}) is then:

$$C_{pp} = C_t \times \left(\frac{1}{n}\right) = \left[F + \left(\frac{1}{M} - \frac{1}{10}\right)T\right] \times n \times \left(\frac{1}{n}\right) = \left[F + \left(\frac{1}{M} - \frac{1}{10}\right)(T)\right]$$

3.3 Method for Developing Survey Cost Model Two

For the approach that assumed a separate FR trip for every GQ in sample, a simulated sample for cluster sizes one through ten was generated in order to accurately express the number of GQs that would be in sample for every cluster size scenario. We again assumed that the number of FRs, the sample size and other administrative costs would remain the same, including *T*. *T* is the same because an FR trip to a sampled cluster is also a trip to a GQ. For this second approach, it was also assumed that the appropriate number of FRs needed to conduct all the cluster sample interviews for a particular GQ facility would be included in the single GQ facility trip. We expressed this as a projected total cost per sample person (C_{pp2}):

$$C_{pp2} = \left[F + \frac{(\text{new GQ facilities in sample}) \times T}{n} \right]$$

As discussed in the introduction, a significant portion of the GQ sample is clustered by month. For the 2009 large GQ sample, roughly 30 percent of sample clusters belonged to GQ types that have their sample clustered by month (the month being randomly assigned). Of the remaining sample, roughly 93 percent was made up of GQs with only one sample cluster. The remaining sample clusters are spread throughout the year. The number of GQ facilities that require multiple FR visits throughout the year is relatively small, so the trip estimator for the second model was seen as appropriate. The ultimate goal was to model the number of FR trips based on both the ACS clustering techniques and the actual number of GQs which receive multiple FR visits. We felt that an appropriate estimate for the number of FR trips would fall somewhere in between models one and two.

3.4 Method for Cost Benefit Analysis

Previous work done on this utilized the idea of an optimization function and defined:

$$G = f(V,C)$$

where V is a variance function and C is a total cost function. G would then be optimized with respect to V and C (Roebuck, 2005). G was then developed into a product of the cost and design effect functions. The design effect was used instead of the entire variance function because of the fixed sample size and rate as discussed in 3.1. Additionally, the product of these two functions would be dependent upon *M* (cluster size), which was the variable of interest in this study. A unit decrease in cluster size would increase the cost function, while simultaneously decreasing the design effect. We examined two separate expressions, G_1 for the first model using sample clusters, and G_2 for the second model using GQs. Each expression is the product of the projected cost per sample person function and the design effect. Finding the minimum of this product gives the largest decrease in variance per increase in cost, with respect to the variable cluster size (Roebuck, 2008).

$$G_1 = C_{pp} \times DE = \left[F + \left(\frac{1}{M} - \frac{1}{10} \right) (T) \right] \times \left[1 + (M-1)\rho \right]$$

$$G_2 = C_{pp2} \times DE = \left[F + \frac{(\text{ new GQ facilities in sample}) \times T}{n} \right] \times \left[1 + (M-1)\rho \right]$$

G_1 allowed for the derivation of an optimal cluster size expression in terms of F, T and ρ. G_1 is at a minimum when:

$$M = \sqrt{\left(\frac{T}{\left(F - T/10\right)}\left(\frac{1-\rho}{\rho}\right)\right)}$$

since:

$$G_1 = F - F(M-1)\rho - \left(\frac{T}{10}\right) - \left(\frac{T}{10}\right)(M-1)\rho + \left(\frac{T}{M}\right) + \left(\frac{T}{M}\right)(M-1)\rho$$
$$= F + FM\rho - F\rho - \left(\frac{T}{10}\right) - \left(\frac{TM\rho}{10}\right) + \left(\frac{T\rho}{10}\right) + \left(\frac{T}{M}\right) + T\rho - \left(\frac{T\rho}{M}\right)$$

We then find the minimum of G_1 with respect to M by setting $\frac{\partial G_1}{\partial M} = 0$ and isolate M as follows:

$$\frac{\partial G_1}{\partial M} = F\rho - \left(\frac{T\rho}{10}\right) - M^{-2}T + M^{-2}T\rho = 0$$

$$M^{-2}(T\rho - T) = \left(\frac{T\rho}{10}\right) - F\rho$$

$$M^2 = \frac{(T\rho - T)}{\left(T\rho/10\right) - F\rho} = \left(\frac{T(\rho-1)}{\left(\left(T/10\right) - F\right)\rho}\right)\left(\frac{-1}{-1}\right) = \frac{T(1-\rho)}{\left(F - \left(T/10\right)\right)\rho}$$

$$M = \sqrt{\left(\frac{T}{\left(F - T/10\right)}\left(\frac{1-\rho}{\rho}\right)\right)}$$

We used this expression to find the optimal cluster size for model one, given 2009 cost data and an estimate of ρ. This was also analyzed using combinations of cost and ρ, and then examining how the optimal cluster size would respond. This allowed for a type of sensitivity analysis for M where we could examine many combinations of cost and ρ without the need to perform ten different calculations for every combination.

Model two involved a simulated GQ sample selection for cluster sizes one through ten. The number of new GQs in sample for each cluster size was then used in the calculation of G_2 (ten values of G_2). The optimal cluster size was then selected for the lowest value of G_2. It is possible to approximate G_2 as an expression with a similar form to G_1 and in turn derive another optimal cluster size expression in F, T and ρ allowing for the same sensitivity analysis (only this time for GQs and not sample clusters). For this study we deemed this unnecessary.

3.5 Simulated GQ Sample Selection for Variable Cluster Size

For the simulation we used the ACS 2009 GQ universe of records and sampled only the large GQs. This was done using the 2009 sampling models with modifications to the cluster size. The number of GQs in sample was then identified for every one of the cluster size simulated samples.

4. Findings

Variance reduction calculations were produced using a ten percent population characteristic, a large GQ sample size of 170,288 from the 2009 GQ sample simulation and a design effect of 1.96. Both methods for finding an optimal cluster size were implemented, also using data from the 2009 simulation. As expected, the optimal cluster size for model one was larger when compared to model two. The final optimal cluster size could then be chosen as the size that fell between these two values.

4.1 Variance Reduction Results

The samples generated by the simulated sample varied only slightly for each cluster size. The estimates for the variance and standard error on that estimate are shown in Table 1 along with the design effect for that cluster size. As expected the rate of variance reduction is fairly linear since we assumed a fixed overall sample size and an intraclass coefficient of 0.1067 for an estimate on a ten percent population characteristic. Therefore the main component affecting change is the design effect, which for this study is a function of the cluster size only.

Table 1. Variance Estimates

Cluster Size	Variance	Standard Error	Design Effect
1	5.2×10^{-7}	7.2×10^{-4}	1.00
2	5.7×10^{-7}	7.6×10^{-4}	1.11
3	6.3×10^{-7}	7.9×10^{-4}	1.21
4	6.8×10^{-7}	8.2×10^{-4}	1.32
5	7.4×10^{-7}	8.6×10^{-4}	1.43
6	7.9×10^{-7}	8.8×10^{-4}	1.53
7	8.5×10^{-7}	9.2×10^{-4}	1.64
8	9.0×10^{-7}	9.5×10^{-4}	1.75
9	9.6×10^{-7}	9.8×10^{-4}	1.85
10	1.0×10^{-6}	1.0×10^{-3}	1.96

4.2 Survey Cost Model Results

The results for the total projected cost functions were based on ACS cost data from 2009. Cost summaries showed a total survey average cost per sample person, which included both large and small GQs. Since the small GQ population makes up a relatively small proportion of the total GQ population (roughly 10 percent), the average cost per sample person was used for F in the cost model calculations. The 2009 GQ universe was made up of 72,844 large GQs and 107,621 small. The small GQ population was 655,471, while the population for the large GQs was 6,635,950.

From ACS cost data we used a total current cost per sample person (*F* from sections 3.2 and 3.3) of $41.35. This was calculated as the total GQ survey cost for fiscal year 2009 divided by the total GQ sample workload for that same time period. For the travel cost per cluster (*T*) we used the 2009 mileage reimbursement amount of $0.55 and the total miles per GQ sample cluster of 55.98 miles (U.S Census Bureau, 2009). We also included the average salary of an FR per GQ of $34.75, and estimated two FRs for every trip (Roebuck, 2008). This resulted in a cost of $100.29 per GQ sample cluster, which was used as the travel cost for both models. Table 2 presents the cost approximations that were used in this study.

Table 2. Cost Approximations

Cluster Size	Projected Cost/Person (model one)	Projected Cost/Person (model two)	Projected Total Cost (model one)	Projected Total Cost (model two)
1	$131.35	$69.25	$22,370,000	$11,790,000
2	$81.35	$60.35	$13,850,000	$10,280,000
3	$64.68	$54.71	$11,010,000	$9,317,000
4	$56.35	$50.79	$9,596,000	$8,649,000
5	$51.35	$48.01	$8,744,000	$8,175,000
6	$48.02	$45.99	$8,177,000	$7,831,000
7	$45.64	$44.42	$7,771,000	$7,563,000
8	$43.85	$43.18	$7,467,000	$7,352,000
9	$42.46	$42.18	$7,231,000	$7,182,000
10	$41.35	$41.35	$7,041,000	$7,041,000

4.3 Results for Cost Benefit Analysis Model One

By using the formula for optimal *M* (section 3.3) along with the data provided in section 4.2, the following result was produced:

$$M = \sqrt{\left(\frac{T}{F - \left(T/10\right)}\left(\frac{1-\rho}{\rho}\right)\right)} = \sqrt{\left(\frac{100}{41 - \left(100/10\right)}\left(\frac{1-0.1067}{0.1067}\right)\right)} = 5.2 \text{ or } \approx \text{cluster size of } 5$$

A range of values for these two variables was desired since these results used an estimate for ρ from 2006 and production data from 2009. A sensitivity analysis was examined since the expression is useful for quick calculations of *M*. We could then examine multiple values of cost and ρ (a limitation is that there are restrictions on *F* and *T*). These values could then be plotted against a desired range of ρ, allowing for an examination of how an optimal cluster size might be influenced by a survey's cost structure and the intraclass correlation coefficient (given the assumptions and limitations of the model). These results are shown in Table 3. There are other limitations, since *T* is dependent upon the current survey cluster size. However, for this study we conducted the analysis given the current survey design and cost structure.

Table 3. Analysis of Travel Cost, ρ and Optimal M Using First Model (Parameters and optimal M for this study are highlighted)

Travel Cost =	$100	$115	$130	$145	$160	$175	$190
ρ = 0.1	5	6	6	7	8	8	9
ρ = 0.2	4	4	4	5	5	5	6
ρ = 0.3	3	3	3	4	4	4	4
ρ = 0.4	2	2	3	3	3	3	4
ρ = 0.5	2	2	2	2	3	3	3
ρ = 0.6	1	2	2	2	2	2	2
ρ = 0.7	1	1	1	2	2	2	2

4.4 Results for Cost Benefit Analysis Model Two

For model two, because an expression for an optimal *M* was not put in a form similar to model one, a cost factor was instead calculated for every cluster size simulation. For every unit decrease in the cluster size, more GQs were introduced into the sample, thereby increasing the projected overall survey cost as expected. An optimal cluster size was then selected. The cost factors can be found in Table 4. The minimum cost factor occurred for cluster size three. This result was not unexpected given the assumptions. Since each GQ facility will receive only one FR visit, travel costs will not be as high as in model one and therefore the cost variance product minimum should occur at a smaller cluster size.

Table 4. Cost Factor Results for Model Two (minimum G_2 is shaded)

Cluster Size	Projected Cost per Case	Design Effect	Cost Factor (G_2)
1	$69.25	1.00	69.25
2	$60.35	1.11	66.79
3	$54.71	1.21	66.39
4	$50.79	1.32	67.04
5	$48.01	1.43	68.49
6	$45.99	1.53	70.52
7	$44.42	1.64	72.85
8	$43.18	1.75	75.42
9	$42.18	1.85	78.18
10	$41.35	1.96	81.06

4.5 Final Estimate of the Optimal Cluster Size

Using the results from the two extremes of our models, the final optimal cluster size could then be selected as the cluster size falling between these two points. The first model which assumed trip growth was equal to the number of clusters in sample resulted in an optimal cluster size of five. The second model assumed trip growth to be equal to the number of GQs in sample and resulted in three as the optimal cluster size. We would then choose four as the optimal cluster size. This idea is presented graphically in Figure 1. Table 5 shows the numbers for both models, showing the growth in the number of GQ facilities in sample and the number of sample clusters as cluster size changes. The estimated number of FR trips to GQ facilities was expected to fall between those two columns. The estimated number of FR trips is presented graphically in Figure 2.

Figure 1. Graphical Representation of Optimal Approximation Technique

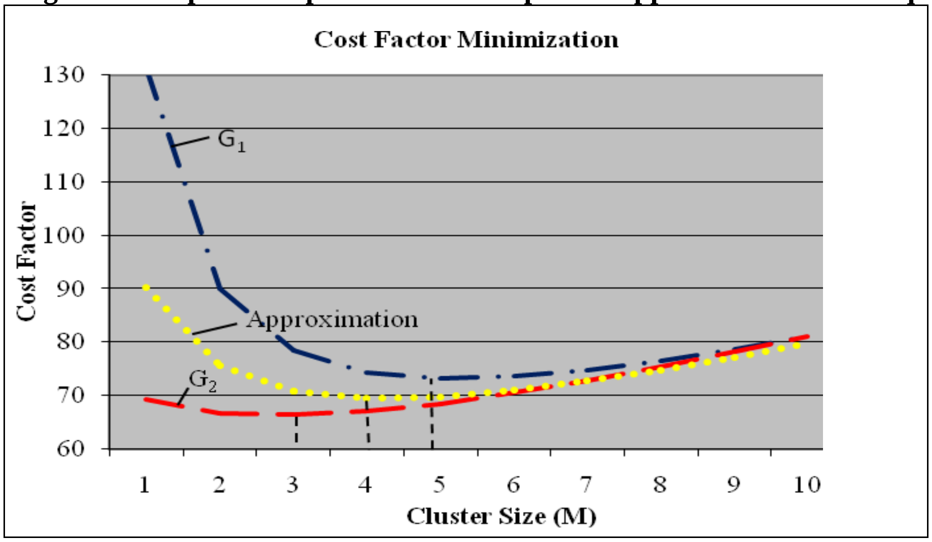
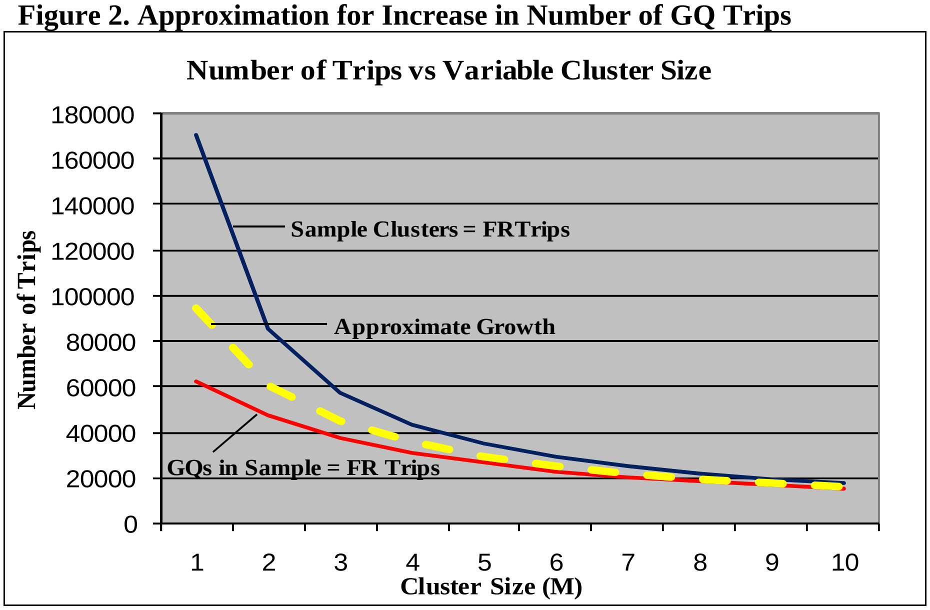


Table 5. Large GQs and Clusters in Simulated Samples

Cluster Size	GQs in Sample	Clusters in Sample	Ratio (Clusters/GQs)
1	62,278	170,288	2.73
2	47,129	85,143	1.81
3	37,521	56,760	1.51
4	30,839	42,569	1.38
5	26,102	34,055	1.30
6	22,662	28,379	1.25
7	19,988	24,329	1.22
8	17,876	21,278	1.19
9	16,178	18,919	1.17
10	14,768	17,027	1.15

Figure 2. Approximation for Increase in Number of GQ Trips



5. Conclusions

Reducing the sample cluster size for large GQs will result in variance reduction while simultaneously increasing the overall travel cost for the survey. The optimal cluster size which gives the largest reduction in variance for the least amount of cost was estimated by first finding the extreme optimal points in terms of assumptions about survey cost growth. Our results indicate the optimal cluster size falling between these two points.

6. Limitations

This study looked only at large GQs, yet the cost data from Field includes both small and large. Since small GQs make up a small enough percentage of the GQ sample (14 percent of GQs and roughly ten percent of expected person level sample) the average total current cost per person was used as the estimate for the large GQ total cost per person.

Another limitation is that the intraclass correlation coefficient is an estimate based on a design factor from the average of the largest design factors in each state from 2006. This estimate must be recalculated using more recent data.

This study also assumes that if the cluster size were to change, FRs would continue to complete approximately 10 interviews per day. So for very large GQs with multiple sample clusters, a single trip to the GQ would include the correct number of FRs needed to complete all the interviews at that GQ (model two).

6. Further Research

Since we have identified several areas in the research where we would like to perform more analysis, and to possibly consider other approaches, this work is considered preliminary research. We would like to find other ways to estimate ρ and use this in our methodology. Additionally, we would like to refine our cost approximation technique, particularly for FR travel cost. We would also like to identify when it would be necessary to increase the number of FRs, since the very small cluster sizes may require additional FRs for the workload. Additionally, we would like to consider fixing the cost or variance and then to examine the optimal cluster size and sample size.

Acknowledgements

We would like to thank Jim Roebuck and Karen King for their work in this study. We are also grateful to Freddie Navarro for his contributions and assistance, and Don Keathley for his useful insights.

References

1. U.S. Census Bureau (2009) "Design and Methodology: American Community Survey" Issued April 2009: http://www.census.gov/acs/www/Downloads/dm1.pdf

2. Cochran, William (1977), *Sampling Techniques 3rd Edition*, New York: John Wiley & Sons

3. 2005 Draft 2, "American Community Survey (ACS): Simulation: Effect of cluster size and intraclass correlation coefficient on variances of group quarters (GQs) estimates"

4. Roebuck, Michael J 2008 Draft 4 "Re-evaluation of Effect of Cluster Size on Cost and Variances of Group Quarters (GQ) Sample"

5. U.S. Census Bureau Internal Memorandum, "ACS/PRCS Group Quarters (GQ) Expenditures through September of Fiscal Year 2009", for All Regional Directors, ACS-GQ Regional Office Memorandum No. 09-24