A NATIONALLY REPRESENTATIVE SAMPLE OF ASIANS THAT IS CITY-BASED

Steven Pedlow, NORC at the University of Chicago

Abstract: The Census Integrated Communications Program Evaluation (CICPE) was designed to evaluate the promotional campaign's effect on Decennial Census participation for six race/ethnicity groups of interest. A nationally representative Core sample was designed to collect interviews for Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites. However, it was impractical to include the rarer Asian, American Indian and Alaska Native (AIAN), and Native Hawaiian and Other Pacific Islander (NHOPI) populations in the Core design. For the Asian sample, we designed a separate area probability sample.

Traditional area probability sampling designs use counties or metropolitan areas as first-stage units, but smaller geographical units can better target hard-to-reach populations. The CICPE Asian sample used cities as the first-stage units. This paper describes a case study on the potential of using smaller geographical units in an area probability design, and reports the challenges of collecting a nationally representative sample for this hard-to-reach population.

Keywords: Area Probability Sampling, CICPE, Decennial Census

Acknowledgement: Lawrence Cahoon of the U.S. Census Bureau helped with the sample design, especially in terms of balancing the increased oversampling/eligibility rate against the increased design effect.

1. Introduction

Every ten years, the U.S. Census Bureau attempts to count every American through the Decennial Census. For the 2000 Decennial Census, the Census Bureau responded to declining mail participation in the 1990 Decennial Census (which increased the costs of in-person enumeration visits) with a greatly expanded outreach and promotion campaign called the "Partnership and Marketing Program" (PMP). NORC at the University of Chicago was contracted to conduct the 2000 Partnership and Marketing Program Evaluation (PMPE), which included a series of three face-to-face in-person surveys: before the PMP; during the PMP; and after the PMP during the non-response follow-up operation of the 2000 Decennial Census. The Census Bureau was sensitive to differential impact of the PMP by race/ethnicity, and so the sample was equally divided among six different race/ethnicity groups, including Asians.

In 2010, the Census Bureau took the lessons learned from 2000 and designed an Integrated Communications Program (ICP) to encourage mail participation in the 2010 Decennial Census. NORC at the University of Chicago again conducted an evaluation of the ICP called the "Census Integrated Communications Program Evaluation" (CICPE) that again utilized in-person face-to-face interviewing. The same six race/ethnicity groups from the 2000 evaluation were again of interest. In 2000 and 2010, NORC's sample designs included a Core sample that was a nationally representative area probability sample to collect interviews from Hispanic, non-Hispanic African-Americans, and non-

Hispanic Whites. However, supplemental samples were necessary for the remaining three race/ethnicity groups, including Asians.

For the 2000 PMPE, NORC's Asian Supplemental Sample collected all interviews from the five U.S. cities with the largest Asian populations. For the 2010 CICPE, we wanted a more nationally-representative sample of Asians. This paper describes how we used a city-based area probability sample to greatly increase the coverage of our Asian sample. Section 2 will discuss the 2000 PMPE and 2010 CICPE sample designs, including a review of Area Probability Sampling. Section 3 will describe the details of the 2010 CICPE Asian sample design. Section 4 will present some results from fielding the 2010 CICPE Asian sample. Section 5 summarizes this paper.

2. The 2000 PMPE and 2010 CICPE sample designs

Both the 2000 PMPE and the 2010 CICPE had sample designs that included three waves of data collection. The first wave of data collection took place before the main campaign elements, the second wave took place while the campaign peaked, and the third wave took place after the mail participation deadline to avoid in-person follow-up had passed. Both designs had a sample size that was an idealized 3,000 interviews per wave divided equally among six race/ethnicity groups: Hispanics, non-Hispanic African-Americans, non-Hispanic Whites, American Indian and Alaska Natives (AIAN), Native Hawaiian and Other Pacific Islanders (NHOPI), and Asians. The second wave of the 2010 CICPE design was compressed into a shorter time period, so the sample size was dropped to 2,100.

A nationally-representative area probability sample called the "Core" sample was designed to collect interviews for the three largest race/ethnicity groups: Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites. National coverage as part of this Core sample was impractical for the three smaller race/ethnicity groups, so three supplemental samples were necessary. This paper focuses on the Asian Supplemental Samples for the 2000 PMPE and 2010 CICPE.

At the time of the 2010 CICPE sample design, the latest source of information on local Asian populations was the 2000 Decennial Census because the American Community Survey had not yet released small area data. According to the 2000 Census, there were 11,898,828 U.S. Non-Hispanic Asians (Barnes and Bennett, 2002), alone or in combination, comprising 4.2 percent of the U.S. population at the time. This figure includes those who marked Asian, regardless of whether other race boxes were marked on the census form; the 2000 Census was the first Decennial Census where race was asked using a "mark all that apply" format.

Since there are many more Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites in the U.S. population than there are Asians, collecting enough Asian interviews through the Core sample would require impractically large screening samples with heavy subsampling of the eligibles for the higher population race/ethnicity groups. In fact, national coverage itself was considered impractical during the planning of the 2000 PMPE, as shown by Wolter et al. (2002), which collected all Asian interviews from the five U.S. cities with the largest Asian populations: New York, Los Angeles, San Francisco, Chicago, and Seattle. At the time of the 2000 Census, 18.8 percent of the U.S. Asian population lived in these five cities. This meant that the 2000 PMPE did not attempt to have a nationally representative Asian sample since the coverage of that sampling frame was only 18.8 of the U.S. Asian population. Within these five cities, 6.5 percent of the population was Asian. If the sample was equal-probability within these cities, the eligibility rate for the Asian sample would be 6.5. We can also refer to this rate as the screening "hit-rate."

For the 2010 CICPE design, our intention was to improve coverage through a national design. Most national face-to-face surveys in the United States use a multi-stage area probability (AP) sampling design that selects clusters of housing units to interview in order to reduce data collection costs (Kish, 1965). In a multi-stage AP sampling design, a set of large clusters are first selected (first-stage units). Within the selected large clusters, sets of small clusters are selected (second-stage units). Finally, within these selected small clusters, individual housing units are selected for interviewing. The basic objective for a multi-stage AP sampling design is a nationally-representative equal-probability sample permitting optimal statistical efficiency. To achieve this, AP samples use probability proportional to size (PPS) sampling in which "larger" areas have a greater selection probability. The measure of size often used for the probabilities is the number of housing units, usually derived from Census data.

Most national area probability samples have first-stage units that are county-based, often using even larger metropolitan statistical areas (MSAs) where present (Lohr, 2009). However, the key idea in this paper is that hard-to-reach populations are better targeted at small geographies. Within the large first-stage unit areas for typical area probability designs, the smaller second-stage areas are often block-based, either in terms of block groups or entire census tracts. Of course, national samples can't use first-stage geographies as small as individual blocks; this would require too many clusters that are too spread out for cost effectiveness. However, NORC has a history of using smaller geographies as first-stage units to better oversample race/ethnicity groups.

The National Longitudinal Survey of Youth 1979 cohort (NLS79) obtained a nationally representative set of interviews with youths who were 14-21 years old while oversampling Hispanic and African-American youths. To do this, NORC split the task into two parts. First, a nationally-representative area probability sample was used to get a nationally-representative mix of Hispanic, non-Hispanic African-American, and non-Hispanic non-African-American youths. A second area probability sample was used to obtain only Hispanic and African-American youths. This "Supplemental" sample did not differ in its design from the nationally representative "Cross-Sectional" sample, but different areas were selected to better target Hispanic and African-American youths.

The National Longitudinal Survey of Youth 1997 cohort (NLSY97) took this design one step further (Moore et al., 2000). In the "Supplemental Sample" design, all first-stage units were counties rather than using entire metropolitan statistical areas in urban areas. Remembering that our goal was to oversample Hispanic and African-American youths, MSAs often have central city counties that have a high rate of minority youths surrounded by outlying, more rural areas that are have lower concentrations of Hispanic and African-American youths. Our strategy allowed us to separate counties with many minority youths from surrounding counties in the same MSA with fewer of them. Counties were still considered too large to target Asians, so we used cites as the first-stage clusters in the 2010 CICPE sample design.

3. The 2010 CICPE Asian Sample

Our first task to select a city-based Asian sample was to construct a sampling frame of cities. The most recent data available at the time was still the 2000 Decennial Census. Our first step was to set a threshold of 1,000 Asians for a city to be included, which led to a set of 1,261 U.S. cities that included 75.6 percent of all U.S. Asians. While our universe did not represent 100 percent coverage, it did comprise a substantial increase over the 18.8 percent coverage for the 2000 PMPE Asian sample design. Within these 1,261 cities, the population is 7.8 percent non-Hispanic Asians. Table 1 gives the coverage and eligibility rate for different non-Hispanic Asian population thresholds that we could have used for our sampling frame of cities:

Table 1. Threshold Options for the Asian Frame of Cities

Minimum Number of Non-Hispanic Asians	**Eligible Cities**	**Asian Population Coverage**	**Eligibility Rate**
100,000	8	20.46%	12.28%
50,000	18	25.92%	12.69%
25,000	46	34.07%	12.30%
10,000	153	48.12%	10.58%
5,000	321	57.87%	9.67%
2,500	653	67.59%	8.72%
1,000	*1,261*	*75.57%*	*7.83%*
500	2,059	80.32%	7.04%
250	3,015	83.17%	6.50%
100	4,569	85.29%	5.97%

Table 1 shows that as the threshold decreases, the coverage increases while the eligibility rate decreases.

Keeping in mind our goal of 500 Asian interviews in each wave, we needed to balance the number of cities where we would have to hire staff against the cluster size determined by the average number of interviews we would need to collect in each city. Increasing the number of cities would increase the cost, while decreasing the number of cities would increase the clustering and therefore the design effect. Balancing these two factors, we decided to select a representative sample of 25 cities (requiring an average of 20 interviews per city) with probability proportional to the city population of non-Hispanic Asians. Our design gave every Asian in our frame of 1,261 cities an equal chance to be in one of our selected cities. New York and Los Angeles, two of the cities used for the 2000 PMPE design, were selected with certainty, while eleven of our twenty-five selected cities were in California. Other states with more than one city selected were Hawaii, New York, and Texas. Table 2 gives summary statistics for our 25 selected cities.

The Asian population sizes in our 25 cities range from the minimum of 1,000 to around 800,000 with a median of approximately 23,000. Asian population percentages range from under 2 percent to over 70 percent with a median of 12.86 percent.

Table 2. Summary Statistics for the 25 Cities in the 2010 CICPE Asian Sample

Statistic	Asian Population	Asian Population Percentage
Minimum	~ 1,000	1.94 percent
10th percentile	~ 3,000	4.46 percent
25th percentile	~ 7,000	5.71 percent
Median	~ 23,000	12.86 percent
75th percentile	~110,000	21.59 percent
90th percentile	~250,000	51.92 percent
Maximum	~800,000	71.22 percent

We then selected entire census tracts as our second-stage clusters within our selected cities. We selected five from each non-certainty city with probability again proportional to the non-Hispanic Asian population, but the certainty cities were properly given six (Los Angeles) or thirteen (New York) clusters, while one city only had two tracts to select (both were selected). Therefore, we selected a total number of 131 census tracts, which resulted in an average of 3.8 interviews per selected Census tract. The selected tracts had an even larger range of Asian population percentages, ranging from 87 percent to less than 1 percent.

An equal probability sample using these 131 census tracts would have resulted in a sample eligibility rate of 7.8 percent (almost twice the national eligibility rate), but this can be increased by oversampling in tracts with a higher proportion of Asian residents. We actually designed a sample with an expected eligibility rate of 26 percent, but this required some tracts to be oversampled by a factor of 50. Differential sampling weights that result from such a skewed oversample would have created a large design effect, which would have greatly reduced our effective sample size.

As in most statistical design decisions, the amount of oversampling involved a balance between lowering screening costs versus keeping variance due to differential weighting low. With help from the Census Bureau, we agreed to limit the design effect so that the loss in effective sample size due to differential sampling would be no greater than 20 percent (a design effect due to differential sampling no greater than 1.25). Our approach decreased the differential oversampling from a factor of 50 to a factor of 3. In so doing, we incurred more screening costs, but maintained the effective sample size closer to the number of interviews. Our specific strategy was to oversample tracts with eligibility rates of at least 20 percent by a factor of 3, and to oversample tracts with eligibility rates between 10 and 20 percent by a factor of 2. With this strategy, our estimated eligibility rate was 12.5 percent, three times as large as the national eligibility rate.

4. 2010 CICPE Asian Sample Field Results

We were able to meet our sample targets for the Asian sample, but our actual unweighted eligibility rates were lower than our estimate for two out of the three waves. It is not surprising when the observed eligibility rate is lower than the planned eligibility rate. Households that are eligible are the most difficult households at which to achieve cooperation (even at the screener level). This means that there is often “hidden non-response” in the screener non-responses. Table 3 shows that the 2000 PMPE achieved higher eligibility rates, which were due to a higher level of oversampling. The 2000 PMPE Asian design oversampled areas with eligibility rates of at least 20 percent by a factor of 5 (Wolter et al, 2002):

Table 3. Planned and Actual Unweighted Eligibility Rates

Statistic	**2000 PMPE**	**2010 CICPE**
First-Stage Eligibility Rate	6.5 percent	7.8 percent
Planned Eligibility Rate	unknown	12.5 percent
Wave 1 Eligibility Rate	22.2 percent	10.3 percent
Wave 2 Eligibility Rate	13.3 percent	12.9 percent
Wave 3 Eligibility Rate	18.9 percent	8.5 percent

Table 4 compares the weighted response rates from 2010 CICPE against the unweighted 2000 PMPE response rates:

Table 4. Response Rates for 2000 PMPE and 2010 CICPE

Wave	**2000 PMPE**	**2010 CICPE**
	(Unweighted)	**(Weighted)**
Wave 1 Response Rate	57.2 percent	50.7 percent
Wave 2 Response Rate	71.0 percent	64.2 percent
Wave 3 Response Rate	60.8 percent	73.8 percent

Weighted rates are not available from the 2000 PMPE, but unweighted rates are not appropriate for the 2010 CICPE because of the mixed-mode data collection procedures that included subsampling of non-respondents for in-person follow-up. Nevertheless, the average response rates for both studies are around 63 percent, and both studies have their lowest response rate in the first wave. The response rates are higher for the 2000 PMPE in the first two waves, but the response rate is much higher in the third wave for the 2010 CICPE.

Non-response bias is usually immeasurable, but the 2010 CICPE study is an exception. With Census Bureau cooperation, we were able to match the entire set of our selected households to 2010 Decennial Census response data. Non-response bias was an important issue for the 2010 CICPE since it was designed around probable/actual response to the 2010 Decennial Census, and it is logical to think that non-respondents to our survey would be more likely to be non-respondents to the 2010 Decennial Census. Table 5 gives the actual mail response rates to the 2010 Decennial Census by April 18 for three different types of 2010 CICPE respondents, as well as interview non-respondents and those households for which we could not determine eligibility (screener non-respondents). The three types of respondents are: 1) Refusers – those respondents who were (soft) refusals at one time, 2) Difficult Respondents – those respondents who had more than the median number of visits before responding, and 3) Easy Respondents – those respondents who responded after less than the median number of visits. All of the mail response rates in Table 5 are weighted. Table 5 shows that our respondents did have higher mail return rates. As expected, the easy Asian respondents had the highest mail return rate by April 18 (64.3 percent) while the non-respondents had the lowest mail return rate by April 18 (53.0 percent). We estimated the proportion of unknown eligibles that would be eligible and counted them as non-respondents. Combining the two non-respondent categories together, the mail return rate was 53.6 percent. Combining the three respondent categories together, the mail return rate was 62.7 percent. Combining all five categories together, the mail return rate for our entire Asian sample was 59.4 percent.

Table 5. Mail Response Rates for the Asian CICPE Sample (Weighted)

Outcome	Return Rate	Response Status	ALL
Unknown Eligibles	54.5 percent	Non-Respondents: 53.6 percent	ALL: 59.4 percent
Non-Respondents	53.0 percent		
Respondents – Refusers	61.9 percent	Respondents: 62.7 percent	
Respondents – Difficult	62.2 percent		
Respondents - Easy	64.3 percent		

The non-response bias is the difference between the estimate for only the respondents (62.7 percent) and the entire population of interest, represented by the entire sample (59.4 percent). Therefore, our Asian sample's non-response bias is 62.7 – 59.4 = 3.3 percent. Since this is positive, our respondents are more likely to be mail responders by April 18 than our non-respondents, which is the expected direction of our non-response bias. Interestingly, of our six race/ethnicity groups, four had almost no bias (less than plus or minus one percent) while our American Indian and Alaska Native respondents were less likely to be mail responders by April 18 than our entire sample of American Indian and Alaska Native housing units. This bias in the unexpected direction may be because our AIAN sample was concentrated on tribal lands, most of which are not eligible for mail return. We used the April 18 date cutoff because this marks the start of the in-person follow-up effort. Any households returning their mail forms after this date may still have been visited in-person.

5. Summary

Even for a hard-to-reach population, it may be possible to attempt a nationally representative sample if the population can be targeted by local areas. Asians make up only 4.2 percent of the U.S. population, but are somewhat clustered by city. We have achieved a 75.6 percent coverage rate for a national sample of Asians for the Census Integrated Communications Program Evaluation (CICPE) study by selecting 25 cities from a frame of 1,261 U.S. cities with a population of at least 1,000 Asians. While we could not come closer to 100 percent coverage in a cost-effective way, and expensive screening was still necessary, we believe that our data is more representative of U.S. Asians than the 2000 PMPE study taking place in only five U.S. cities as well as any list-assisted telephone survey using Asian surnames or other methods with unknown biases.

6. References

Barnes, J. and C. Bennett (2002). "The Asian Population: 2000, a Census 2000 Brief." U.S. Census Bureau. Available online at:
http://www.census.gov/prod/2002pubs/c2kbr01-16.pdf

Datta, R., T. Yan, D. Evans, S. Pedlow, B. Spencer, and R. Bautista (2012). "The 2010 Census Integrated Communications Program Evaluation (CICPE) Final Report." U.S. Census Bureau. Available online at:
http://2010.census.gov/2010census/pdf/2010%20Census%20Integrated%20Communications%20Program%20Evaluation.pdf

Kish, L. (1965). Survey Sampling. Wiley: New York.

Lohr, S. (2009). Sampling: Design and Analysis, Second Edition. Duxbury Press: Pacific Grove, California.

Moore, W., S. Pedlow, P. Krishnamurty, and K. Wolter (2000). "The National Longitudinal Survey of Youth 1997 (NLSY97) Technical Sampling Report." NORC at the University of Chicago. Available online at:
http://www.nlsinfo.org/preview.php?filename=nlsy97techsamprpt.pdf

Wolter, K., B. Calder, E. Malthouse, S. Murphy, S. Pedlow, and J. Porras (2002). "Census 2000 Evaluation: Partnership and Marketing Program Evaluation," U.S. Census Bureau Planning, Research and Evaluation Division: Washington, D.C. Available online at: http://www.census.gov/pred/www/rpts/D.1.PDF