THE PLANNING DATABASE: Its Development and Use as an Effective Targeting Tool in Census 2000

by Antonio Bruce and J. Gregory Robinson
Population Division
U.S. Census Bureau

For presentation at the Annual Meeting of the Southern Demographic Association, Arlington, VA, October 23-25, 2003

This paper reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

USCENSUSBUREAU
Helping You Make Informed Decisions

Introduction

The planning database for Census 2000 assembled a range of housing, demographic, and socioeconomic variables that are correlated with nonresponse and undercounting. The database provided a systematic way to identify potentially difficult-to-enumerate areas to flag for special attention in Census 2000.

The variables included in the planning database (PDB) were guided by extensive research conducted by the Census Bureau and others to measure the undercount and to identify reasons people are missed (de la Puente, 1995). These variables include housing indicators (percent renters, multiunits, crowded housing, lack of telephones, vacancy) and person indicators (poverty, not high school graduate, unemployed, complex household, mobility, language isolation). Other operational and demographic data were included (such as nonresponse rates and race/ethnic distributions). Using the 1990 census as the initial source, a database containing these variables was developed for all tracts in the country for use in the planning, implementation, and evaluation of Census 2000. (U.S. Bureau of the Census, 1999). The PDB contains "hard-to-count" (HTC) scores which summarize the attributes of each tract or block group in terms of enumeration difficulty.

The predictive effectiveness of the planning database and HTC scores was proven by testing against empirical measures of nonresponse and net undercount in the 1990 census, 1995 test census, and the Census 2000 Dress Rehearsal (Robinson and Kobilarcik, 1995; Robinson, 1996; Word, 1997; Bruce and Robinson, 2001). In this paper, we demonstrate the effectiveness of the PDB as a targeting tool in Census 2000 with specific examples. We show how (1) the PDB was used to identify hard-to-count areas and determine the relative size of these areas, (2) the PDB with 1990 data was an excellent predictor of mail response rates in Census 2000, and (3) the PDB was effectively used to target areas with concentrations of populations that speak a language other than English. In preparing for Census 2010 and preceding census tests, we can capitalize on the planning database's targeting power and descriptive statistics for small areas.

Development of the Planning Database for Tracts

As noted, the planning database assembled a range of housing, demographic, and socioeconomic variables that are correlated with nonresponse and undercounting. The database provides a systematic way to identify potentially difficult to enumerate areas. Using the detailed data from the 1990 census, a database containing the variables below was developed for all tracts in the country. In addition, the PDB contains "hard-to-count" (HTC) scores which summarize the attributes of each tract or block group in terms of enumeration difficulty. The database layout and description of each variable is as follows:

Variable	Description	Used in HTC score
Gidtract	State/County/Tract Code	
State	FIPS State Code	
County	FIPS County Code	
Tract	Census Tract Code	
Total POP	Total Population -100%	
Total HU	Total Housing Unit (HU) -100%	
HTC	Hard-to-Count Score	
Pct Vacant	Vacant HU's	X
Pct 10+ Multi Unit	Multi-Unit: 10+ in structure	
Pct 2+ Multi Unit	Multi-Unit: 2+ in structure	X
Pct Renter	Renter-occupied unit	X
Pct Crowded	Units with more than one person per room	X
Pct No H/W HH	Not Husband/Wife Household (HH)	X
Pct HU no Phone	HU without a telephone	X
Pct Not HS Grad.	Not High School Grad (no Diploma)	X
Pct Poverty	Persons below poverty level	X
Pct Pub. Assist	Receiving public assistance income	X
Pct Unempl.	Unemployed	X
Pct Ling Iso HH	Linguistically Isolated Household	X
Pct Move 89-90	Householder moved in unit 1989 or 1990	X
Pct Black	Black or African American	
Pct Am. Indian	American Indian/Aleut/Eskimo	
Pct API	Asian or Pacific Islander	
Pct Hispanic	Hispanic Origin	
Pct NonrespR	1990 Non Response Rate	

The 1990 file contained 58,405 records, one for each tract.
Applying the same methodology used in the 1990 PDB, the Census 2000 database was developed that includes the variables in the planning database (PDB)

Development of the Hard to Count Scores

The PDB file contains "hard-to-count" (HTC) scores which summarize the attributes of each tract in terms of enumeration difficulty. A total of 12 variables correlated with nonresponding households and undercounting were used to derive the HTC score (see variables marked with an 'X' in previous section that describes the variables).

A set of algorithms to determine HTC scores was used as follows:

(1) each individual variable was sorted across geographic areas from high to low (e.g., sort tracts from highest percent poverty to lowest),

(2) scores (0 to 11) were assigned to each variable for each tract (e.g., values of 11 were given to tracts with the highest poverty rates of over 44.3 percent and values of 0 were given to tracts below the national poverty median of 9.9 percent in 2000),

(2) the scores assigned to each of the 12 variables for a tract were summed to form a composite HTC score for the tract.

Table 1 illustrates the HTC scores and percentile distribution of tracts in the 2000 census for three specific variables: percent renter, percent not husband/wife household and percent poverty.

With twelve variables used to produce the HTC scores in the tract file, the scores can range from 0 to 132. The comparative standing of areas provides indicators of the likely degree of difficulty in enumeration. Areas with the highest scores (e.g., over 60) are likely to be the areas with relatively high nonreturn rates and undercount while areas with the lowest scores are likely to be areas with low rates. Table 3 (in later section) summarizes the distribution of tracts on the hard-to-count continuum in 1990 and 2000 and illustrates the strong association of HTC scores and nonreturn rates (a correlation coefficient of 0.77 statistically demonstrates the association over all tracts in 2000):

Table 1. Percentile Distribution of Hard-to-Count (HTC) Variables for Tracts: 2000

Census 2000		Range of Percentile Distribution		
Range or Point	**HTC Score**	**% Renter Occupied Unit**	**% Not Husband Wife HHs**	**% Persons below Poverty**
97.5 -100	11	91.3 - 100	83.9 - 99.2	44.3 - 100
95 - 97.5	10	82.3 - 91.3	78.8 - 83.9	37.2 - 44.3
90 - 95	9	69.8 - 82.3	72.0 -78.8	29.3 - 37.2
85 - 90	8	60.9 - 69.8	66.9 - 72.0	24.3 - 29.3
80 - 85	7	53.7 - 60.9	62.9 - 66.9	20.6 - 24.3
75 - 80	6	47.8 - 53.7	59.4 - 62.9	18.0 - 20.6
70 - 75	5	42.9 - 47.8	56.4 - 59.4	15.9 - 18.0
65 - 70	4	38.5 - 42.9	53.6 - 56.4	14.0 - 15.9
60 - 65	3	34.5 - 38.5	51.2 - 53.6	12.5 - 14.0
55 - 60	2	31.3 - 34.5	49.0 - 51.2	11.1 - 12.5

50 - 55	1	28.2 - 31.3	46.8 - 49.0	9.9 - 11.1
< 50	0	< 28.2	< 46.8	< 9.9
Note: See text for description of HTC algorithms to assign HTC scores.				

Uses of the Planning Database in Census 2000

1. Provided systematic basis to profile areas on a "hard-to-count" continuum.

The detailed housing, demographic, and socioeconomic variables in the PDB provided a systematic way to profile areas in terms of potential ease or difficulty of enumeration. The "hard-to-count" (HTC) score was used as a key summary statistic on attributes of each tract in terms of enumeration difficulty.

Table 2 shows the relative distribution of neighborhoods on a HTC continuum and their socioeconomic profiles–specific to type of race or Hispanic concentration:

Tracts with African-American, Hispanic, and American Indian/Eskimo/Aleut (AIEA) majorities are disproportionately located in hard-to-count (HTC) neighborhoods. Over three-quarters of American Indian majority tracts (118 tracts, or 77.6 percent), over two-thirds of Hispanic majority tracts (3,031, or 73.1 percent), and over one-half of African-American majority tracts (3,655, or 60.6 percent) are concentrated in the HTC category (defined as tracts with scores of 60+). In contrast, less than 1 percent of tracts with high concentrations of Non-Hispanic Whites (184 tracts) fall in this category.

As a whole, tracts with Asian and Pacific Islander majorities are not concentrated in HTC areas–the tract distribution is spread over all categories (60+, 30-59, <30). Underlying PDB data show that the HTC attributes vary widely by ethnic group (e.g. Vietnamese, Thai, and Cambodians tend to have higher scores than Chinese, Japanese, or Filipinos).

The planning database provides empirical data on why these neighborhoods are difficult to enumerate. Compared to the national average (see last row in Table 2), the tracts with relatively high HTC scores exhibit concentrations of renters, crowded units, complex households, poverty and other variables associated with hard-to-enumerate conditions. These characteristics hold for every race/ethnic category in the high HTC category shown in Table 2. The high HTC tracts with concentrations of Hispanics or Asians also show high levels of linguistic isolation. Thus it is not surprising that the mail return rate is relatively low in these tracts (col. 3)--the low rate is predictable given the characteristics of these neighborhoods (see more discussion in next section).

In Census 2000, the planning database and HTC scores provided another source to alert the field staff to potentially challenging enumeration areas and the composition of their populations. Computer generated maps provided a valuable tool to profile these areas and study the geographic clustering of hard-to-count neighborhoods.

Table 2. Distribution of Tracts in Hard-to-Count (HTC) Categories and Associated Attributes by Race/Origin: 2000 Census

(for all groups except Non-Hispanic Whites, the universe includes only tracts where the race/origin group represents 50% or more of the total tract population. For Non-Hispanic Whites, the universe includes only tracts where Non-Hispanic Whites represent 90% or more of the total tract population)

				Selected "HTC" Variables (Percent)				
Race/Origin and HTC Category	Number of Tracts	Pct. Distr	Mail Return Rate	Renter Units	Crowded Units	No Phone	In Poverty	Linguistic Isolation
African-American	**6,031**	**100.0**	**64.3**	**49.5**	**3.6**	**5.5**	**25.9**	**2.7**
60+	3,655	60.6	59.9	62.5	4.8	7.5	33.5	3.6
30-59	1,729	28.7	67.7	37.5	2.5	3.7	19.3	1.8
<30	647	10.7	75.6	21.4	1.2	1.1	8.5	1.2
Hispanic/Origin	**4,150**	**100.0**	**68.6**	**53.5**	**16.4**	**4.7**	**25.3**	**22.7**
60+	3,031	73.1	66.1	62.8	18.8	5.8	29.6	25.5
30-59	985	23.7	74.1	32.3	11.5	2.4	15.4	15.7
<30	134	3.2	78.5	19.3	4.3	1.0	8.1	9.5
Amer. Indian	**153**	**100.0**	**62.4**	**31.2**	**13.1**	**28.0**	**35.9**	**10.0**
60+	118	77.6	61.9	25.0	16.2	34.0	38.3	12.5
30-59	33	21.1	62.7	33.3	2.5	7.9	23.6	1.8
<30	2	1.6	-	-	-	-	-	-
Asian/Pacific Islander	**406**	**100.0**	**75.3**	**45.1**	**11.5**	**1.4**	**11.8**	**19.6**
60+	126	31.0	69.0	75.9	19.5	3.1	23.5	32.4
30-59	115	28.3	73.7	43.9	12.1	1.2	10.6	18.0
<30	165	40.6	81.5	21.1	4.5	0.3	5.0	10.4
Non-Hisp White	**21,188**	**100.0**	**81.4**	**21.1**	**0.4**	**1.8**	**7.7**	**0.9**
60+	184	0.9	69.8	60.1	1.5	5.5	28.8	4.5
30-59	2,869	13.5	75.4	33.1	0.7	4.6	17.1	1.1
<30	18,135	85.6	82.4	18.9	0.3	1.3	6.2	0.8
All Tracts	**62,599**	**100.0**	**76.1**	**33.9**	**2.7**	**2.4**	**12.3**	**4.2**

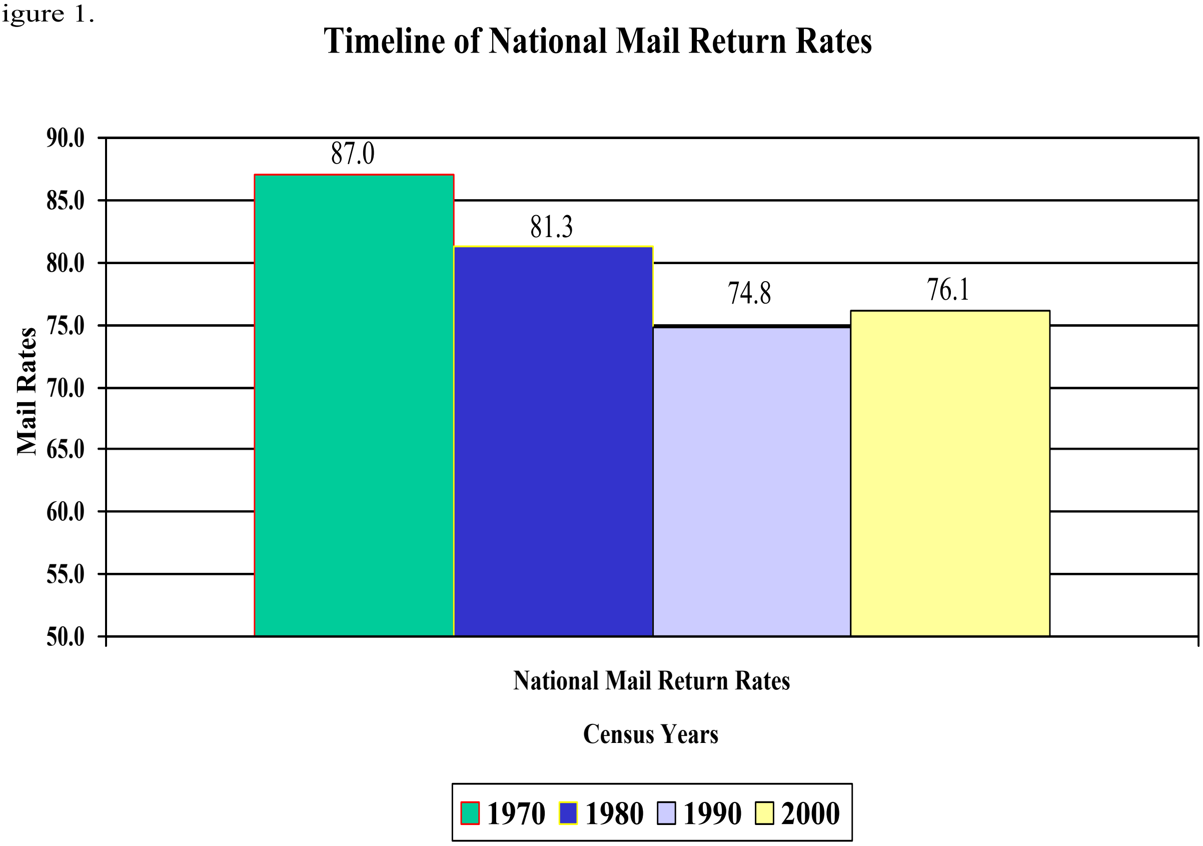
See text for description of Hard-to-Count Scores.

The Asian category includes Native Hawaiian and Pacific Islanders; the American Indian category includes Alaskan Natives. The race groups were based on the reporting of "race alone" in response to the race question in Census 2000.

Note on HTC variables:
(1) Crowded variable is defined as % of units with more than one person per room.
(2) No Phone variable denotes telephone service is not available in unit
(3) Linguistically isolated variable represents the % of linguistically isolated households among all households in those tracts.

2. Predicted Mail Return Rates in 2000 based on 1990 Mail Rates and Hard-to-Count

Figure 1.



Attributes

The national mail return rate dropped dramatically from 1970 to 1990, falling from 87.0 to 74.8 (see figure 1). A goal in Census 2000 was to reverse this downward trend, which was achieved as the return rate for 2000 edged up to 76.1 percent.

The 1990 Planning Database and associated HTC scores proved to be a powerful predictor of patterns in mail response and mail return rates in Census 2000. Using the 1990 PDB, we assigned all tracts in the country into 10 mutually exclusive strata based on Hard-to-Count (HTC) Scores. The deciles (see Table 3) span the spectrum of response rates ranging from very low response rates in areas with concentrations of hard-to-count attributes to very high response rates in areas with an absence of hard-to-count characteristics (note the inverse relationship). In using the 1990-based HTC scores for the 2000 analysis, we assume that the demographic/ socioeconomic/ housing makeup of an entire stratum in 2000 is essentially the same as in 1990.

We compared patterns of response rates according to HTC scores. The 1990 and 2000 response rates shown in Table 3 and displayed in Figure 2 for tracts classified by HTC score are remarkably similar. The response rates vary systematically along the HTC continuum. The Census 2000 return rate was 61.7 percent in 2000 (58.3 in 1990) for the decile of 5,815 tracts with highest concentrations of hard-to-count attributes (HTC scores of 76+); the Census 2000 return rate was a much higher 85.4 percent (84.8 in 1990) in the decile stratum with the lowest concentrations (HTC scores less than 2).

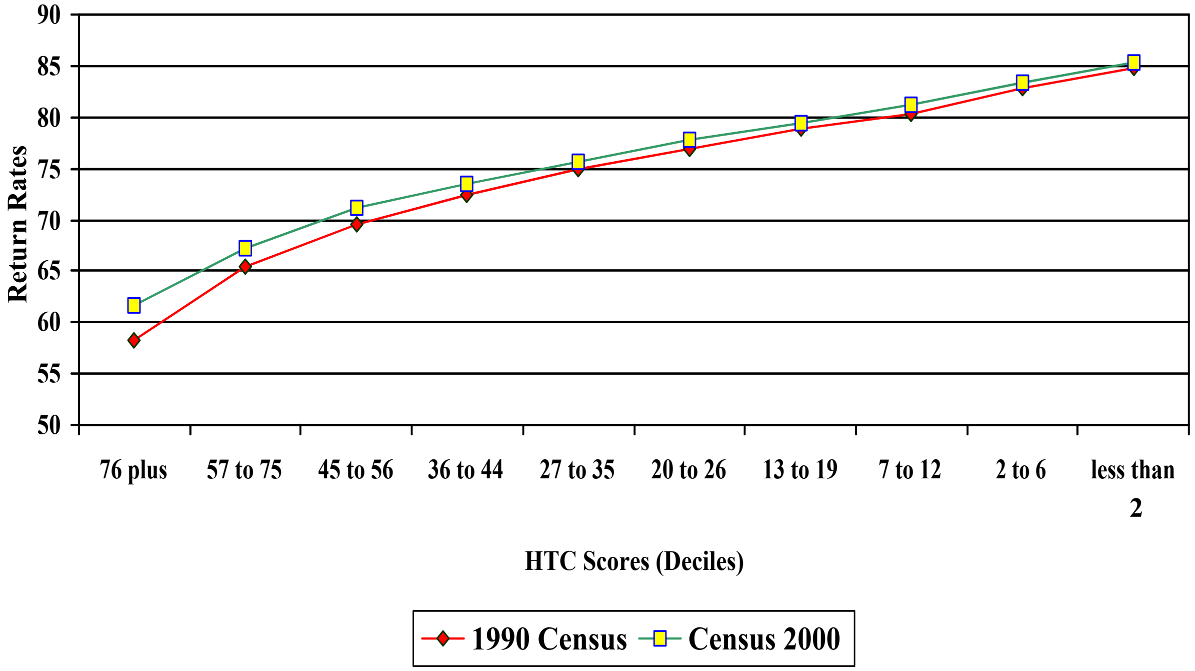
Despite the uniformity of response patterns by HTC decile, differentials are observed in the increase in rates from 1990 to 2000. The response rates rose by the largest amount (by 3.4 percentage points) in the most difficult-to-enumerate areas (Strata 1). The second greatest gain (1.8 points) was in the second most difficult strata. The lowest response rate increases are observed in the "easier-to-enumerate" deciles (strata 9 and 10; with slight increases of 0.6 points).

Table 3. Comparison of 1990 and Census 2000 Mail Return Rates by Hard-to-Count Strata

Hard-to-Count Scores	No. of Tracts	1990 Mail Return Rates	No. of Tracts	2000 Mail Return Rates	Mail Return Change, 1990 to 2000
76 plus	**5,815**	58.3	**6,349**	61.7	**3.4**
57 to 75	**6,077**	65.4	**6,994**	67.2	1.8
45 to 56	**5,762**	69.5	**6,125**	71.1	1.6
36 to 44	**5,504**	72.5	**5,493**	73.5	1.0
27 to 35	**6,391**	74.9	**6,203**	75.7	0.8
20 to 26	**5,476**	76.8	**5,468**	77.8	1.0
13 to 19	**6,039**	78.4	**6,259**	79.5	1.1
7 to 12	**6,033**	80.3	**6,784**	81.3	1.0
2 to 6	**6,326**	82.8	**7,085**	83.4	0.6
< 2	**4,982**	84.8	**5,839**	85.4	0.6
Total	**58,405**	**74.8**	**62,599**	**76.1**	**1.3**

Figure 2.

Comparison of Trends in 1990 and 2000 Mail Return Rates by Hard-To-Count Scores



3. Targeted Areas with Concentration of Linguistically Isolated Households.

We used the variables in the planning database to identify linguistically isolated (LI HH's) households and languages based on 1990 and 2000 census data for tracts in each State. In this example, we focus on tracts where 15 percent of the LI HH's spoke a language other than English, that is, tracts or neighborhoods where a relatively high degree of households had language difficulty and may need special targeting.

Table 4 provides evidence of the geographic concentration of LI areas at the State level. In 2000, 4,944 tracts in the nation had linguistically isolated household concentrations of 15 percent or more, and 83 percent of these tracts were located in just 6 States (California, New York, Texas, Illinois, Florida, and New Jersey). California alone accounted for one-third of the LI tracts in 2000. The geographic patterning was similar in 1990.

Table 4. Top Ranked States of Linguistically Isolated Households[1]

1990 Census				Census 2000			
Rank (1990)	No. Tracts	Pct. of Tracts	No. of Tracts: 60 plus HTC	Rank (2000)	No. Tracts	Pct. of Tracts	No. of Tracts: 60 plus HTC
U.S.	**3,004**	**100.0**	**2,482**	**U.S.**	**4,944**	**100.0**	3,760
1 Ca	915	30.5	742	1 Ca	1,670	33.8	1,261
2 NY	696	23.2	552	2 NY	897	18.1	658
3 TX	463	15.4	423	3 TX	708	14.3	564
4 IL	181	6.0	157	4 IL	296	6.0	198
5 NJ	166	5.5	130	5 FL	278	4.7	169
6 Fl	149	5.0	96	6 NJ	234	5.6	175
Six States	2,570	85.6	2,100	Six States	4,083	82.6	3,025
Other States	434	14.1	382	Other States	861	17.4	735

The hard-to-count scores in the PDB can be used to examine if LI populations tend to live in areas that exhibit characteristics associated with difficulty of enumeration. The association is strong. In both 1990 and 2000, about three-quarters of LI tracts had hard-to-count scores of 60 or more (3,760 of 4,944 tracts in 2000) and fit the profile of high HTC tracts shown in Table 2.

The PDB allows us to systematically identify the concentrations of linguistically isolated (LI) households by type of language. Table 5 illustrates the top 5 ranked LI languages (by number of tracts) in selected States in 1990 and Table 6 shows the rankings for a more limited grouping of languages in 2000. In both 1990 and 2000, linguistically isolated Spanish household predominate by far in every State. The 2nd to 5th ranked languages vary considerably by State. So not only are linguistically isolated populations geographically concentrated, but the specific language involved has a particular geographic pattern as well.

Table 5. Top 5 Ranked Linguistically Isolated Languages by State[2] : 1990

[1] LI languages are based on a threshold where at least 15% of households in the tract are linguistically isolated. Source: 1990 Census Planning Database and Census 2000 SFT3.

[2] LI languages are based on a threshold where at least 15% of households in the tract are

	CALIFORNIA		NEW YORK		TEXAS	
	LI	No.	LI	No.	LI	No.
	Lang.	Tracts	Lang.	Tracts	Lang.	Tracts
Rank 1	Spanish	842	Spanish	584	Spanish	442
Rank 2	Chinese	132	Chinese	147	Chinese	3
Rank 3	Korean	48	Italian	81	Cambodian	3
Rank 4	Vietnamese	47	Russian	71	Vietnamese	3
Rank 5	Cambodian	40	Korean	37	Korean	2

	ILLINOIS		NEW JERSEY		FLORIDA	
	LI	No.	LI	No.	LI	No
	Lang.	Tracts	Lang.	Tracts	Lang.	Tracts
Rank 1	Spanish	162	Spanish	157	Spanish	138
Rank 2	Polish	40	Portuguese	29	Creole	30
Rank 3	Chinese	10	Polish	15	French	10
Rank 4	Korean	8	Italian	9	Vietnamese	1
Rank 5	Ukranian	4	Korean	7	N/A	-

Table 6. Top 5 Ranked Linguistically Isolated Languages by State[3] : 2000

	CALIFORNIA		TEXAS		NEW YORK	
	LI	No.	LI	No.	LI	No.
	Lang.	Tracts	Lang.	Tracts	Lang.	Tracts
Rank 1	Spanish	1,090	Spanish	644	Spanish	379
Rank 2	API	223	API	3	Indo-Euro	140
Rank 3	Indo-Euro	25	Indo-Euro	-	API	76
Rank 4	Other	-	Other	-	Other	-

	FLORIDA		ILLINOIS		NEW JERSEY	
	LI	No.	LI	No.	LI	No.
	Lang.	Tracts	Lang.	Tracts	Lang.	Tracts
Rank 1	Spanish	213	Spanish	168	Spanish	139
Rank 2	Indo-Euro	17	Indo-Euro	21	Indo-Euro	15
Rank 3	API	2	API	11	API	5
Rank 4	Other	-	Other	-	Other	-

Discussion

linguistically isolated. Only 32 languages were included in this analysis. Source: 1990 Census

[3] LI languages are based on a threshold where at least 15% of households in the tract are linguistically isolated. Only 4 cluster of languages were included in this analysis. Source: Census 2000 Planning Database

Using 1990 and 2000 tract-level data, this paper demonstrates how the Planning Database and associated Hard-to-Count Scores were highly effective in targeting potentially difficult-to-enumerate areas. The PDB flagged areas that experienced low mail response rates in Census 2000 and identified areas with concentrations of linguistically isolated households.

While the PDB clearly has potential use in the 2010 census, we need to identify applications for the PDB to aid the planning/evaluation of ongoing current surveys and test censuses that lead up to the decennial census. In addition, we need to develop ways to update the PDB and HTC scores by incorporating data more recent that 2000, such as inclusion of results of the American Community Survey and administrative data.

References

Bruce, Antonio, J. Gregory Robinson, and Monique V. Sanders. 2001. "Hard-to-Count Scores and Broad Demographic Groups Associated with Patterns of Response Rates in Census 2000", Proceedings of the Social Statistics Section, American Statistical Association.

de la Puente, Manuel. 1993. "Why Are People Missed or Erroneously Included by the Census: A Summary of Ethnographic Coverage Reports", Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations. Richmond, VA.

Robinson, J. Gregory, and Edward L. Kobilarcik. 1995. "Identifying Differential Undercounts at Local Geographic Levels: A Targeting Database Approach"' paper presented at the Annual Meetings of the Population Association of America, San Francisco, April 8, 1995

Robinson, J. Gregory. 1996. "Demographic Review of the Housing and Population Results of the 1995 Test Census", Memorandum for Arthur J. Norton, March 12, 1996.

Word, David L. 1997. "Who Responds/Who Doesn't? Analyzing Variation in Mail Response Rates During the 1990 Census"' Population Division Working Paper Series, No. 19, U.S. Bureau of the Census.

U.S. Bureau of the Census. 1999. 1990 Data for Census 2000 Planning, CD-ROM with documentation, Washington, D.C.