The sampling bias in random digit dialing

by A. Dianne Schmidley[1]
Bell Atlantic Corporation

All research is based on information from data collection efforts, whether the data result from the use of qualitative approaches, such as, content analysis, focus group sessions, or in-depth interviewing by psychologists, social workers or ethnographers, or from methods which produce data more amenable to quantitative analyses, such as, sample surveys and censuses (technically, a census is a 100% sample survey). A study[2], conducted by Derek Phillips in the early 1970s, concerning the preference of sociologists for either quantitative or qualitative data collection approaches, revealed that more than 90 percent of the research conducted by those social scientists resulted from the analysis of data collected through the administration of interview schedules and/or questionnaires developed for use in a survey setting.

While sociologists are generally viewed as major collectors and users of the data resulting from surveys, they are not the only individuals interested in survey data. Pollsters, advertisers, program administrators and others have important uses for survey data. In fact, there are several broad categories of surveys including:

1. Attitude and public opinion polls,
2. Marketing research, including advertising and public relations surveys,
3. Government surveys, especially those conducted for the purposes of developing legislation and administering and evaluating the effectiveness of programs,
4. Special surveys conducted by researchers in university settings, usually the basis of primary research, and often government funded,
5. Other surveys, including internal

[1]Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Marina Del Rey, California on May 21-24, 1986

[2]Knowledge from What?

organizational surveys conducted to monitor attitudes and opinions.

Individuals with training in survey research methods direct many of these collection activities; however, many surveys, particularly those in the areas of marketing, public relations and opinion research, are conducted by individuals who have no formal training and are not aware of the many pitfalls of survey data. Once the ill-gotten information is translated into copy, it acquires a life of its own and the 'facts' become almost impossible to erase.

The use of data from a survey, for some purpose other than that for which it was collected, is common. For example, data from government surveys such as the Current Population Survey (CPS) are often used as the basis of marketing decisions, while data from a public opinion survey may influence the decisions of government policy makers.

The data from all five categories of surveys are utilized to support policy decisions affecting the expenditure of millions of dollars. Given the penchant of policy makers for basing decisions on data collected through the use of surveys, the need for a critical appraisal of each of the various stages of the survey process has evolved. Questionnaire/interviewer schedule design, sample selection, the administration of the survey process, the collection, tabulation and interpretation of the data, and testing of the reliability and validity of the information collected from the survey, compared to some benchmark, have all become subareas of the survey research process and are carefully scrutinized by trained social scientists.

The purpose of this paper is to shed additional light on the problem of selecting a representative sample of a population to be surveyed, using the procedure of simple random sampling.

When a sample is selected for a survey, attention must be given to the parameters of the universe the researcher hopes to study, and the method of eliciting information from that universe. In the case where the universe to be studied is a human population, there are a limited number of ways of operationally defining the universe and collecting information from the individuals comprising that universe. Each type of definition will affect the method of sample selection chosen. If the universe is operationally defined by street address records, the researcher can mail a questionnaire to each individual respondent. If telephone numbers constitute the universe, the researcher can telephone respondents and administer a questionnaire/interview schedule. With street address records, the researcher can meet with each respondent and conduct interviews face to face.

Generally, evaluations of each of these three approaches suggest that the higher the quality of the data collected (given proper development of the survey instrument), the greater is the expense of the survey. Mail-back questionnaires are the least expensive way in which to collect information and result in the poorest quality data with the lowest response rates, while in-person interviews are the most expensive way in which to collect the information, but produce better respondent cooperation. Given this set of circumstances most researchers employ the telephone as a means of collecting survey information, since it produces medium quality data at medium cost.

I mentioned earlier that when drawing a sample from a universe, the researcher operationalizes the sampling process by giving the members of the populaton s/he hopes to study concreteness in the form of a street address or a telephone number. Client/customer specific lists, collected by some agencies and firms, can be used when the researcher is interested in some subset of the population, and thereby has access to the specific names, telephone numbers and/or

address records of that group.

In most instances, however, the universe to be studied is only vaguely known. It may be comprised of all the individuals in some specific geographic location such as a trading area, county or school zone, or it may be a subset of the population, such as households with a certain income, or people of certain ages or ethnicity. At any rate, owing to a lack of specific information which could be used to contact and interview the individuals of interest in the population, the researcher more often uses a method of simple random sampling (SRS) of available street records or telephone numbers in order to delineate those units which will be sampled for the purpose of collecting information which can then be ascribed to the larger universe from which the sample was drawn.

The preference of survey researchers for telephone interviewing as a means of collecting information from respondents, coupled with the need to employ SRS in order to to identify the individuals to be surveyed within the universe in which they are located, has led recently to the development of a technique called 'Random Digit Dialing' (RDD).

RDD is believed by the naive to be a cure for drawing biased or non-representative samples from a universe. Misinformed advocates of RDD persist in believing that universal telephone service means that everyone has a telephone in their home. Secure in this belief, RDD practicioners promote the idea that by simply dialing the telephone in a random manner one can draw a random sample of the population of any given geographic area. Census data, which are used by experienced survey researchers, to design sampling frames and to calibrate survey results, indicate that universal telephone service does not mean that a telephone is found in every housing unit.

'Universal telephone service' is akin to the notion of 'full employment'. Full employment is defined by many as the situation in which about 5% of the population in the labor force, ages 16 through 64, and desiring employment, is unemployed. There are many persons between the ages of 16 and 64 who are not employed, and are not seeking employment; so full employment does not mean that 5% of this entire age group is unemployed. Similarly, universal telephone service does not mean that every housing unit contains telephone network access. There are a variety of reasons why a unit may be without access to the telephone network.

Attachment A to this paper gives the reader the sources utilized to determine something called telephone penetration. As can be seen in this attachment, there are basically four ways in which to measure telephone penetration: (1) using decennial census data, (2) using information from the CPS, (3) using GBF-DIME records, and (4) comparing household counts with counts of telephone access lines in given geographic areas.

Examples of the decennial census and CPS measures of telephone penetration of each of jurisdictions served by Bell Atlantic appear in Table 1, entitled, "Telephone Availability For Selected Areas: United States and Bell Atlantic Served States, 1980-1985". Figures 1 and 2 illustrate changes over time in telephone penetration in these geographic locations. Table 2 provides and indication of the type of statistic that can be developed using the Access Lines/Households ratio.

Several conclusions are possible from a careful examination of these various measures of telephone penetration:

1. Data from different questions result in different statistics; i.e., the two censuses produced different results partly owing to question wording. These differences,

although found in a comparison of data from the 1970 and 1980 censuses, are illustrated in Table 1 in the 'unit' and 'availability' measures from the CPS,

2. In addition to these differences, however, there are major differences between geographic locations (Figures 1 and 2) and,

3. Differences have occurred over time (Table 2),

4. Differences result from the sample selection process. Table 1a reveals the differences in sample drawn from 1980 census data (Columns a and b); Table 1b illustrates the range associated with each confidence interval computed for the various point estimates of penetration computed from CPS data.

Thus far, we have examined the penetration rate for the total number of households. When differences in types of households is taken into account, the variation in telephone penetration becomes even greater. The data in in Table 3 are from a 1980 Census Public Use Microdata Sample (PUMS). Table 3 contains a cross-tabulation of households by age of householder, household income, and telephone penetration for the state of Virginia. These data indicate that the presence of a telephone in the housing unit is not a random event, but rather that one is less likely to have a phone if one is poor and/or young.

Figure 3 illustrates the ratio of telephone access lines to households in the Bell Atlantic region from the early 1950s to the 1990s forecast period. A surprising event occurred in the late 1970s: the number of access lines increased to the point of outnumbering households. How can this be? The answer is quite simple, it's called multiple lines. In the state of New Jersey alone, more than 10% of households have at least two lines, and these households are not poor and they are not young.

Can one draw a random sample, utilizing RDD? Yes, but a random sample of what? Certainly one will not draw a random sample of the households in a given geographic location. One will draw a random sample of telephone lines and, as we have seen, these are not evenly distributed across the population.

Does this matter? It depends on what one is trying to accomplish. A pollster, trying to predict an election, will want to examine age specific voting patterns and calibrate these to the universe of households with telephones, allowing for multiple lines, of course. Since the people who vote tend to be those with higher incomes, and voter participation increases with age, you will reach the group you wish to sample using telephone interviewing, although you may overstate the case, especially if you are tracking Republican candidates. (Table 4)

A marketing researcher, must be careful to calibrate the resulting data with census data, because the upscale population will be over-represented and other groups under-represented in the sample. An academic or government researcher, conducting a survey designed to produce information for the development of legislation or the operation of a social program, would need to calibrate the data with census or CPS data in order to ensure a representativeness.

In conclusion, it seems highly unlikely that telephone interviewing is a reasonable replacement for the Bureau of the Census traditional decennial census methods of data collection employing a housing address list, mailout questionnaires and enumerators. Given the problems associated with the universe of telephone access lines, if telephone interviewing becomes the mainstay of data collection for the U.S. Government, we will have lost one of the most important means we have of analysing the composition of our population. ¤

Attachment A

METHODS OF MEASURING TELEPHONE AVAILABILITY/PENETRATION (STATE LEVEL)

I. Decennial Census of Population and Housing

Questionnaire

1970 #H1 Is there a telephone on which people living in your quarters can be called?

☐ yes what is the number?

\- - - - - - - - - - -
Phone number

☐ no

1980 #H26 Do you have a telephone in your living quarters?

☐ yes ☐ no

Caveats:
- Timliness of the data
- Validity & Reliability issues

II. Current Population Survey

Interview

Supplementary questions regarding telephone availability asked in March, July & November of each year of a state based sample of households

Caveats:
- Sampling error at state level
- Coverage of non-Bell Atlantic served areas

III. Geobased Files/Dual Independent Map Encoding

Matching telephone customer address records with street address records in the GBF

Caveats:
- Non-urbanized areas not covered
- New housing not covered, although it can be inferred from telephone company records

IV. Residence Acess Lines/Household Estimates;

Use telephone company access line information in the numerator and household estimates prepared by company demographer in the denominator

Caveats:
- Second line development; failure of business office to identify second line
- Accuracy of household estimates; OK for internal purposes, may not stand up in legal/regulatory setting

Table 1-A

TELEPHONE AVAILABILITY FOR SLECTED AREAS: UNITED STATES AND BELL ATLANTIC SERVED STATES, 1980-1985

GEOGRAPHIC UNIT	1980 (A)	1980 (B)	NOV 1983		MARCH 1984		JULY 1984		NOV 1984		MARCH 1985		JULY 1985		6 MONTH AV. (1-6-85)	
			Unit	Avail.	Unit	Avail.	Unit	Avail.	Unit	Avail.	Unit	Avail.	Unit	Avail.	Unit	Avail.
United States	92.9	92.9	91.4	93.7	91.8	93.6	91.6	93.8	91.4	93.6	91.8	93.7	91.8	93.9	91.6	93.7
Washington D.C. (C&P)	96.0	95.3	94.7	95.6	96.1	97.5	93.5	95.4	95.1	96.0	91.6	93.5	93.6	94.9	94.1	95.5
Maryland (C&P of MD)	96.0	95.8	96.3	96.7	96.1	96.9	94.9	95.7	96.1	96.8	95.2	96.2	96.2	98.1	95.8	96.7
Virginia (C&P of VA)	92.6	91.8	93.1	94.7	93.2	95.1	93.0	95.6	92.9	94.6	92.8	94.5	90.4	92.3	92.6	94.5
West Virginia (C&P of WVA)	89.6	89.3	88.1	91.1	87.2	93.5	86.5	90.0	89.4	92.1	88.1	91.4	88.7	92.8	88.0	91.8
New Jersey (NJ Bell Tel.)	95.4	95.3	94.1	95.1	93.5	95.0	96.0	96.9	94.5	96.3	95.1	96.5	95.4	96.5	94.8	96.1
Pennsylvania (Bell of PA)	96.0	95.9	95.1	97.1	94.4	96.0	95.1	96.4	95.1	97.2	94.2	95.5	95.8	96.8	94.9	96.5
Delaware (Diamond State)	95.2	95.1	95.0	96.6	95.4	96.3	93.7	95.1	93.7	95.8	96.6	97.4	94.4	96.1	94.8	96.2

Sources: 1980(A): PHC 80-S1-1 "Provisional Estimates of Social, Economic and Housing Characteristics", Table #1; 1980. (Official numbers files by AT&T with the DOJ, RE. The Modified Final Judgement, 1982.)

1980(B): STF3 - Final Numbers from the 1980 Census of Population & Housing.

Nov. 1983 - July 1985: Current Population Survey (CPS)

Table 1-B

Confidence Intervals Associated With Point Estimates
Of Telephone Penetration for Bell Atlantic Served Jurisdictions

Prepared by D. Schmidley
Demographic Studies
x8638

(1) State	(2) Unit Measure Nov. 1985	(3) CVx	(4) SEx	(5) Range 68%	(6) Range 95+ %
DC	95.6	0.0149	1.42	94.2 - 97.0	92.8 - 98.4
DEL	93.4	0.0119	1.11	92.3 - 94.5	91.2 - 95.6
MD	95.3	0.0120	1.14	94.2 - 96.4	93.0 - 97.6
NJ	94.1	0.0093	0.88	93.2 - 95.0	92.3 - 95.9
PA	95.8	0.0059	0.57	95.2 - 96.4	94.7 - 96.9
VA	92.0	0.0156	1.44	90.6 - 93.4	89.1 - 94.9
WVA	86.1	0.0183	1.58	84.5 - 87.7	82.9 - 89.3

Formula from U.S. Bureau of Census

$$SE_x = X \cdot CV_x$$

X1 = 68% Confidence Interval

X2 = 95% Confidence Interval

SE_x = Standard error

X = Penetration rate

CV_x = Coefficient of variation

Table 1-C

This table is from the U.S. Bureau of the Census - Current Population Survey

TABLE 1 C

PERCENTAGE OF HOUSEHOLDS WITH A TELEPHONE BY HOUSEHOLDER'S AGE

	ALL RACES		WHITE		BLACK		HISPANIC ORIG.	
	Unit	Avail	Unit	Avail	Unit	Avail	Unit	Avail
8 MONTH AVERAGE								
TOTAL HOUSEHOLDS	91.7	93.8	92.3	95.0	80.4	84.8	81.1	84.3
16-24 YRS OLD	77.4	83.6	80.0	85.6	57.9	69.7	62.8	74.1
25-54 YRS OLD	91.8	93.8	93.5	95.1	80.2	84.5	82.7	85.4
55-59 YRS OLD	94.9	96.1	96.0	96.9	87.2	89.6	87.4	89.7
60-64 YRS OLD	95.1	96.0	96.0	96.8	87.9	89.8	88.8	90.4
65-69 YRS OLD	95.9	96.7	96.9	97.5	87.9	90.2	90.4	91.9
70-99 YRS OLD	95.5	96.6	96.1	97.2	89.1	91.3	86.8	90.1
NOVEMBER 83								
TOTAL HOUSEHOLDS	91.4	93.7	93.1	95.0	78.8	83.9	80.7	84.6
16-24 YRS OLD	76.6	84.1	80.2	86.2	49.9	68.2	64.9	71.9
25-54 YRS OLD	91.5	93.7	93.4	95.2	78.7	83.3	81.8	85.6
55-59 YRS OLD	95.0	96.1	96.1	97.0	86.3	88.5	89.3	89.3
60-64 YRS OLD	95.5	96.4	96.4	97.2	89.5	90.7	87.3	90.2
65-69 YRS OLD	95.5	96.2	96.5	97.0	87.2	89.0	90.7	90.7
70-99 YRS OLD	95.4	96.5	96.0	97.0	90.1	92.3	85.5	89.1
MARCH 84								
TOTAL HOUSEHOLDS	91.8	93.6	93.3	94.9	80.1	84.1	80.7	83.6
16-24 YRS OLD	77.8	84.0	80.3	85.5	57.9	71.5	59.0	66.2
25-54 YRS OLD	91.9	93.7	93.5	95.0	80.4	84.0	83.2	85.6
55-59 YRS OLD	94.9	95.9	95.7	96.6	87.6	89.9	88.7	90.5
60-64 YRS OLD	94.2	95.3	95.9	96.7	81.7	85.0	87.4	89.6
65-69 YRS OLD	96.1	96.6	97.0	97.4	87.8	89.3	85.8	87.8
70-99 YRS OLD	95.3	96.3	96.2	97.1	87.2	88.8	82.2	85.5
JULY 84								
TOTAL HOUSEHOLDS	91.6	93.8	93.2	95.0	80.5	85.3	81.1	84.6
16-24 YRS OLD	77.0	83.3	79.4	85.3	60.4	70.0	62.9	70.8
25-54 YRS OLD	91.7	93.8	93.4	95.1	79.8	84.9	83.1	85.8
55-59 YRS OLD	95.1	96.3	96.1	97.1	87.5	90.2	87.4	91.4
60-64 YRS OLD	95.0	96.2	95.8	96.9	87.7	89.5	88.1	90.5
65-69 YRS OLD	96.4	97.1	97.3	97.9	89.3	91.3	88.7	90.6
70-99 YRS OLD	95.2	96.5	95.9	96.9	89.6	93.1	84.0	88.5
NOVEMBER 84								
TOTAL HOUSEHOLDS	91.4	93.6	93.1	95.0	78.9	84.0	81.1	84.5
16-24 YRS OLD	76.1	83.4	79.0	85.4	56.3	70.8	60.8	70.8
25-54 YRS OLD	91.4	93.6	93.3	95.1	78.5	83.3	83.1	85.8
55-59 YRS OLD	94.9	96.2	96.3	97.5	84.7	87.4	85.3	88.3
60-64 YRS OLD	95.6	96.5	96.5	97.3	90.3	92.1	86.0	87.2
65-69 YRS OLD	96.0	96.7	97.1	97.6	86.7	89.1	96.2	96.2
70-99 YRS OLD	95.3	96.6	96.1	97.2	88.0	90.7	87.1	88.8
MARCH 85								
TOTAL HOUSEHOLDS	91.8	93.7	93.3	95.0	80.1	84.4	81.2	84.1
16-24 YRS OLD	77.3	83.1	79.6	84.8	59.8	70.0	62.4	67.1
25-54 YRS OLD	91.9	93.8	93.6	95.2	79.5	83.9	83.0	85.5
55-59 YRS OLD	94.9	95.9	95.8	96.7	87.3	89.1	86.5	89.1

Table 2

% OF HOUSEHOLDS WITH TELEPHONES
IN THE C&P SERVED AREA OF VIRGINIA

Year (1)	Residence Access Lines-In Service (June 30) (2)	Bell Served Households (July 1) (3)	Telephone Penetration (2)/(3)X100 (4)
1960	564,824	763,000	74.0
1961	588,783	782,000	75.3
1962	616,381	802,000	76.9
1963	640,932	823,000	77.9
1964	669,764	843,000	79.5
1965	702,113	863,000	81.4
1966	736,798	882,000	83.5
1967	776,020	904,000	85.8
1968	819,205	933,000	87.8
1969	860,518	975,000	88.3
1970	896,711	1,011,000	88.7
1971	933,814	1,027,000	90.9
1972	972,268	1,068,000	91.0
1973	1,025,160	1,104,000	92.9
1974	1,069,475	1,134,000	94.3
1975	1,097,747	1,168,000	94.0
1976	1,132,060	1,202,000	94.2
1977	1,166,408	1,225,000	95.2
1978	1,212,158	1,263,000	96.0
1979	1,255,582	1,305,000	96.2
1980	1,288,870	1,355,000	95.1
1981	1,319,998	1,386,000	95.2
1982	1,343,961	1,416,000	94.9

Source: (1) Data for Residence Access Lines-In-Service are taken from the Virginia monthly No.7 report, years 1960-80, and taken from the Virginia No. 2705 report for years 1981-82;

(2) Household data is based on household counts from the Decennial Census 1960, 1970 and 1980; and are based on the Current Population Survey and P-25 Population Reports in other years. These reports originate at the U.S. Bureau of the Census.

October 1983

Table 3

Telephone Penetration by Age and Income, Virginia, 1980

	16 to 24	25 to 29	30 to 34	35 to 39	40 to 44	45 to 49	50 to 54	55 to 59	60 to 64	65 to 69	70 to 74	75 to 79
	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER	HSHLDER
Household Income												
Less Than $10,000	0.68	0.75	0.77	0.79	0.78	0.76	0.80	0.83	0.86	0.89	0.92	0.91
$10,000 - 14999	0.79	0.87	0.88	0.88	0.89	0.90	0.90	0.94.	0.95	0.97	0.97	0.97
$15000 - 19999	0.83	0.92	0.94	0.94	0.94	0.94	0.95	0.97	0.96	0.97	0.98	0.95
$20000 - 24999	0.92	0.96	0.97	0.97	0.96	0.97	0.96	0.98	0.98	0.97	0.99	0.99
$25000 - 29999	0.95	0.98	0.98	0.99	0.98	0.99	0.97	0.99	0.99	0.98	0.99	0.98
30000 - 34999	0.91	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.98
$35000 - 39999	1.00	0.98	0.98	1.00	0.98	0.99	0.99	0.99	1.00	0.99	1.00	0.97
$40000 - 44999	0.94	0.97	1.00	0.99	1.00	1.00	0.99	0.98	0.99	1.00	1.00	1.00
$45000 - 49999	0.95	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.98	1.00	0.97	1.00
$50000 or More	0.98	0.98	1.00	0.99	0.99	0.99	0.99	1.00	0.99	0.98	0.99	0.97
Total	0.78	0.89	0.92	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.93	0.93

These data were generated from the 1980 Public Use Microdata Sample by the Pennsylvania State Data Center, at the request of A.D. Schmidley, Bell Atlantic

Table 4

TELEPHONE PENETRATION
BY AGE OF HOUSEHOLDER (%)
(1980 CENSUS)

AGE	RATE	AGE	RATE
18-24	83.8	50-54	96.0
25-29	92.0	55-59	96.2
30-34	94.5	60-64	96.2
35-39	95.3	65-69	96.0
40-44	95.5	70-74	96.2
45-49	95.8	75+	95.6

This data was generated from the Public Use Microdata Samples PUMS for the Bell Atlantic served jurisdictions of Washington, D.C., Delaware, Maryland, New Jersy, Pennsylvania, Virginia and West Virginia, combined.

Figure 1

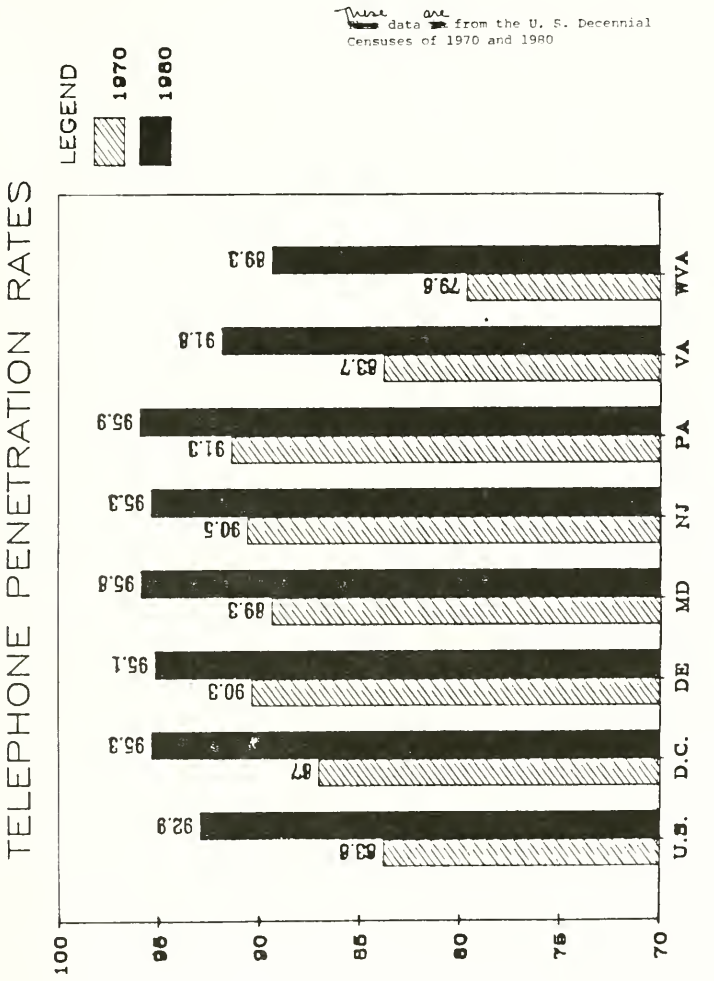

These data are from the U. S. Decennial Censuses of 1970 and 1980
LEGEND
1970
1980
TELEPHONE PENETRATION RATES
89.3
79.6
WVA
91.8
83.7
VA
95.9
91.3
PA
95.3
90.5
NJ
95.8
89.3
MD
95.1
90.3
DE
95.3
87
D.C.
92.9
83.8
U.S.
100
95
90
85
80
75
70

Figure 2

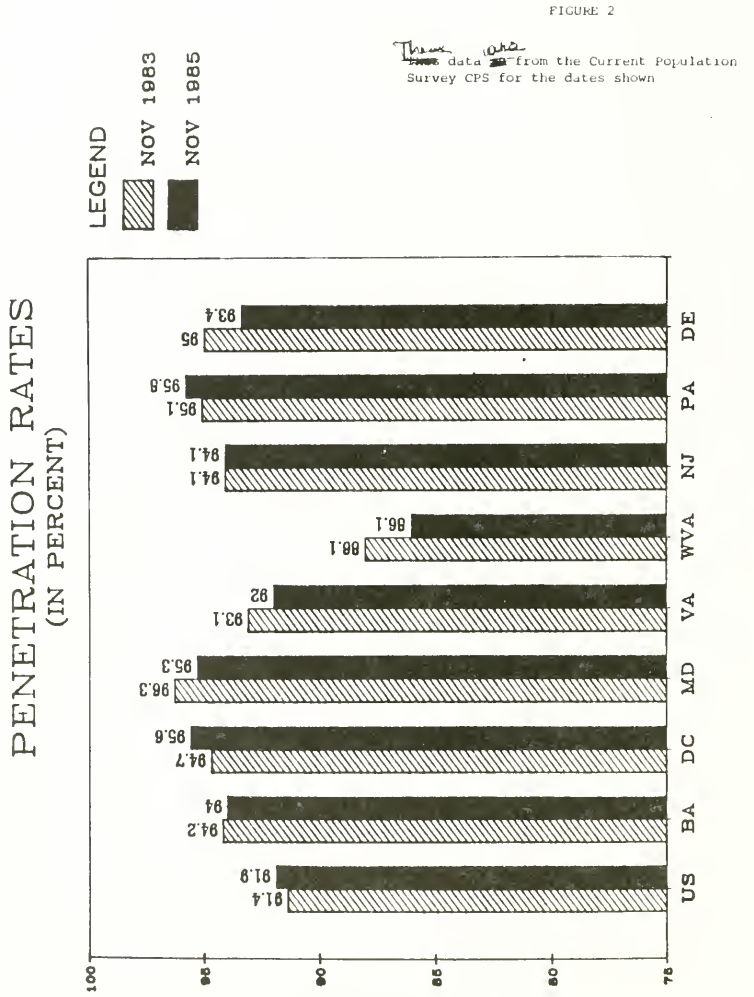
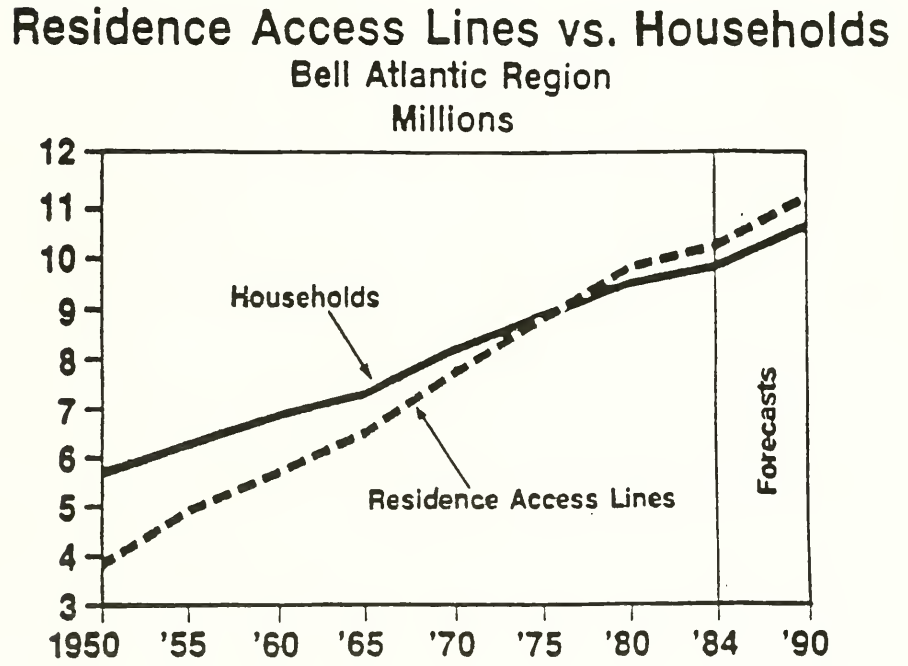


FIGURE 2
data from the Current Population Survey CPS for the dates shown
LEGEND
NOV 1983
NOV 1985
PENETRATION RATES
(IN PERCENT)
US 91.4 91.9
BA 94.2 94
DC 94.7 95.6
MD 96.3 95.3
VA 93.1 92
WVA 86.1 86.1
NJ 94.1 94.1
PA 95.1 95.8
DE 95 93.4
100
95
90
85
80
75

Figure 2


Residence Access Lines vs. Households
Bell Atlantic Region
Millions
12
11
10
9
8
7
6
5
4
3
Households
Residence Access Lines
Forecasts
1950
'55
'60
'65
'70
'75
'80
'84
'90