

USING ADMINISTRATIVE LISTS TO SUPPLEMENT COVERAGE
IN HARD-TO-COUNT AREAS OF THE POST-ENUMERATION SURVEY
FOR THE 1988 CENSUS OF ST. LOUIS

Glenn Wolfgang, Bureau of the Census*
Washington, D.C. 20233

KEY WORDS: Census, Post-Enumeration Survey, Undercount, Response Correlation Bias

I. INTRODUCTION

The Administrative List Supplement (ALS) to the 1988 Post-Enumeration Survey (PES) is a research project testing the feasibility of using administrative lists to improve the coverage of a coverage measurement tool. The PES measured census coverage - whether, for a particular area or group, the census counted too few persons (producing an undercount) or too many persons (producing an overcount). Coverage measurement is important because, as demographic analysis has shown, certain groups have been undercounted and at a rate quite different from other groups. Improvement of that coverage measurement is important because, as demographic analysis has also shown, the PES estimates are biased where too many of the same persons are missed jointly in both the census and the PES. The result of this lack of independence is called response correlation bias.

Demographic analysis uses birth, death, and migration records to track and project population for age, race, and sex groups in large areas. Demographic analysis results, like results from the 1980 Post-Enumeration Program (PEP), a precursor to the PES, have shown that there is more undercount for black males than for whites or females (Fay, Passel, and Robinson, 1988, p.83). However, the PEP estimates of undercount were up to 16.6% lower than demographic analysis estimates - especially for young adult and middle-aged black males. Fay et al. attribute that to correlation bias.

Sex ratios, usually expressed as the number of males per 100 females, have also been used to evaluate coverage measurement estimates. Demographic analysis produces sex ratios that are so stable within age and race groups that they are viewed as standards for comparison. Black males aged 20 to 44 are 95.7% to 97.6% (Fay et al., p. 84) as numerous as black females of the same age. Yet corresponding ratios for almost all 1980 PEP data were 6% to 13% lower. Again, Fay et al. point to correlation bias, saying, "Whatever factors make males hard to count in the census appear to be at work in the PEP also."

The ALS used administrative lists (records from several state and federal agencies) to identify persons who should have been, but were not, counted in the original PES interview. The search was limited to a target group, males aged 20-44 and living in minority renter blocks of the sample, because reports such as Fay et al. show that group is hard to count and that its estimates need the most

improvement. Followup interviews verified those who belonged in a supplement which could be included in PES processing and estimation.

The value of the ALS is in improving the accuracy of PES estimates of undercount where there is evidence of response correlation bias. Census coverage may be accurately estimated without complete PES coverage, unless there is response correlation bias. Complete PES coverage would eliminate this bias. Increasing PES coverage with sources more independent of the census than the PES would reduce an existing correlation bias. The ALS aims at more complete PES coverage and reduced correlation bias for the target population, those most likely to be missed by both sources. This report evaluates the contribution of ALS to these ends.

II. THEORY AND BACKGROUND

Coverage measurement with a post-enumeration survey is based on dual system estimation theory. Marks, Seltzer, and Krotki (1974) and Wolter (1986) provide detail beyond the following brief exposition, especially with regard to assumptions on which the estimates depend. There are many such assumptions. Many of them point to ways counts could be incorrectly assigned into the table or to cells within the table. The ALS was designed to deal with the independence assumption which has been a major concern in PES estimation.

The census is one system for estimating population; the PES is a second. The dual system estimate of the true population size, N , for an area or group is based upon crossclassifying them as in Figure 1. Each person in the real population was either listed or missed in each system. The characteristics of the persons listed are compared from one system to the other and, when they match, the person is assigned to the x_{11} cell. The x_{12} and x_{21} cells represent the unmatched residuals. Those missing in both systems (in x_{22}) cannot be observed.

	In PES	PES Missed	All
In Census	x_{11}	x_{12}	x_{1+}
Census Missed	x_{21}	x_{22}	x_{2+}
All	x_{+1}	x_{+2}	N

Figure 1. The Dual System Estimation Scheme

The two systems are independent if knowledge of a person's status in one system does not tell us more than we otherwise knew about status in the other. Independence may be expressed mathematically by proportionality across levels of one system (or the other). For example, independence implies that the expected proportion of those enumerated by the census is the same among those in the PES as in the general population:

$$E(x_{11}/x_{+1}) = E(x_{1+}/N). \quad (1)$$

This expression of independence was used by Marks et al. It is related to and could be derived from Phi, the correlation coefficient formulae for a two-by-two contingency table, set to equal zero or from the cross-product ratio, a measure of association attributed to Goodman, set to unity (Marascuilo and McSweeney, 1977, pp 207-208). The cell values in each case should be non-zero expected values or probabilities for individuals' assignment to the cell. Wolter defines independence between sources using the cross-product ratio.

After replacing expected values in (1) with observed values and rearranging, the independence statement yields an estimator of the total population based upon the census total and the PES total and the count of matches:

$$\hat{N} = x_{+1}x_{1+}/x_{11}. \quad (2)$$

Estimation based on formula (2) therefore depends on an assumption of independence. If the unobserved x_{22} cell were disproportionately large, it means that being missed by the census goes along to some degree with being missed by the PES. This is response correlation, which biases the dual system estimator. In this case, the bias is downward, underestimating both the population and the undercount.

Because the x_{22} cell count never becomes known through the census or PES, other sources must be used to assess and ameliorate response correlation bias. Demographic analysis produces alternative population estimates for evaluating PES results. Administrative lists are used here to improve the data, i.e. lessen the possibility or extent of correlation bias.

	In PES	PES Missed	All
In Census	$x_{11} + x_{13}$	$x_{12} - x_{13}$	x_{1+}
Census Missed	$x_{21} + x_{23}$	$x_{22} - x_{23}$	x_{2+}
All	$x_{+1} + x_{+3}$	$x_{+2} - x_{+3}$	N

Figure 2. Dual System Estimation With Supplement (ALS) Data

When additional persons found by a supplementary source (namely, administrative lists) are included in dual system estimation as PES cases, their tallies are in effect removed from the "PES missed" and added to the "in PES" cells of the crossclassification. Figure 2 demonstrates this. The terms with subscripts 1 and 2 refer to the same counts as in Figure 1. Supplement person counts have a subscript 3.

The dual system estimator employing the supplement data then becomes:

$$\hat{N}^s = (x_{+1} + x_{+3})x_{1+}/(x_{11} + x_{13}). \quad (3)$$

In order to improve upon the estimate of N when it is underestimated due to correlation bias, as Fay et al. have shown it to be for the ALS target population, the estimate based on supplement data should be greater:

$$\hat{N}^s > N. \quad (4)$$

Algebraic substitution from formulae (2) and (3) above expresses this in terms of observable match rates, the proportions matched to the census, within the separated PES and supplement cases:

$$x_{11}/x_{+1} > x_{13}/x_{+3}. \quad (5)$$

If, for a population shown to have downward correlation bias, the PES match rate (on the left side of the inequality) is smaller than the supplement match rate (on the right), the supplement must be subject to an even more severe correlation bias than the PES and will only make the estimate worse.

III. PROCEDURES

The Administrative List Supplement was conducted in conjunction with the Dress Rehearsal for the Bicentennial Census and its Post-Enumeration Survey. The ALS sample was limited to the St. Louis site and, more specifically, to those PES sample blocks identified as having primarily black renter residents.

Names, addresses, birthdate, race, sex, and social security number of St. Louis residents were requested from several state and federal sources. State agencies provided two lists of those seeking Employment Security services and the driver's license file, the largest procured for the ALS. Federal agencies supplied Selective Service registrants and those on Veteran's Administration rolls. Names and Addresses from the Internal Revenue Service, augmented by characteristics from the Social Security Administration, comprised another list. All files were extracted from their source files by the summer of 1987.**

Each list was read onto mass computer storage, standardized into a common format, geocoded (census geography code numbers assigned to each record on the basis of its address), and culled to select sufficiently complete records relevant only to the target group. The selections were accomplished by keeping records with (1) both first and last name,

(2) a geocodeable address, (3) a ZIP Code found in sample blocks when doing selections prior to geocoding, (4) a geocode of a sample block, (5) male gender, and (6) age between 20 and 44.

The lists were merged, sorted by sample block, and printed onto listings for clerical use. Since many names came from more than one source, clerks tagged duplicated names for omission and transcribed useful data onto the record that was kept. They checked each address to make sure it belonged in the sample. In large blocks, this meant excluding addresses not in the PES subsample segments. A major step in identifying supplement persons was eliminating those already listed in the PES. ALS clerical work coincided with other PES clerical processing.

Supplement names, addresses, and characteristics were clerically matched to those in the census. Each supplement person was confirmed in a field followup interview to be a resident in the sample area on the dates for census and PES enumerations. Supplement data sometimes needed clarification for other reasons. When a supplement-census match seemed likely but still uncertain, followup tried to resolve it. Addresses which did not match exactly to a PES address but seemed to belong in the sample were visited and a map was sketched for use in verifying its existence and location. The main ALS timing concern was having PES matching and followup form preparation done in ALS blocks in time for supplement persons to be checked against and integrated with the PES field caseloads. There was some sampling of supplement cases sent to followup. The reasons and means of the sampling are explained in the next section.

IV. SAMPLE DESIGN

The sample for the Administrative List Supplement is largely determined by that of the Post-Enumeration Survey. The PES sample consists of all persons listed in the original PES interview as residents of the PES sample areas at the time of the interview (July, 1988). The supplement includes everyone listed in available administrative records and confirmed by a followup interview to be a resident of specified areas at the time of the PES but not listed by the PES. In other words, the supplement is anyone in specific areas and groups who can be found who should have been but was not included in the PES.

PES sample area units are blocks demarcated by roadways and waterways. Before selection, all census blocks were assigned to strata designed to be as homogeneous as possible on characteristics such as region of the country, rural-urban location, race/ethnicity of residents, and whether residents own or rent their home. Blocks are selected within each stratum.

If a PES sample block contained 70 or more housing units, the block was subsampled in order to reduce field workloads to 35-50 addresses per block. The subsampling was done by dividing the block into equal clusters and randomly selecting one of the clusters.

The ALS concentrated in the hardest-to-count PES stratum. Although the Dress Rehearsal PES was conducted in areas of the state of Washington and rural Missouri, the supplement was developed only in one of four strata in the city of St. Louis. That stratum's residents are primarily black renters.

Supplement cases, but not the original PES persons, were subsampled at the time of followup in order to save on field costs. The result is some loss of precision in statistics generated using the supplement research data. But it was an opportunity to test a strategy for optimizing the productivity of field followup of supplement persons. The number of supplement followup cases was several times more than projected. Due to the age (vintage) of the information on the administrative records, it is likely that many in the supplement followup would have moved before Census Day or the PES and are therefore out of scope. The rate of moving in the target population is likely much higher than in the general population. The older the record, the higher the rate of moving. The followup sampling strategy adopted was designed to minimize the number of times field interviewers found themselves asking about persons who were long gone from the address.

Followup cases were selected within two strata. Persons judged most likely to end up resolved and still residing in sample were followed up with certainty. A person in that certainty stratum was distinguished by having a record vintage no earlier than 1986, by having had at least one duplicate record omitted, and by sharing a surname with a PES person at that address. Also, all supplement persons matched or possibly matched to a census person were put in this certainty group. Any supplement person with a missing or old record vintage or no duplicates or no surname match was systematically sampled for followup at a one in four rate from a list ordered by name alphabetically within block.

For every sampling step there are weights assigned to sample persons. One weight is the inverse of the selection probability within a stratum. Another weight corresponds to subsample selection. Supplement persons have a followup selection weight. The tabulations resulting from all the processing are multiplied by the product of these sampling weights in estimating statistics representative of the whole population. In this paper only followup selection weights have been used.

Another stratification system, called poststratification, was designed for accuracy and precision of estimation. The PES area sample was drawn before data collection; the poststrata are identified after processing. Age categories and gender are used as well as region, urban-rural, race/ethnicity, and owner-renter. The supplement was restricted to males in two age categories whose population was most underestimated by past post-enumeration programs relative to demographic analysis. It is important to keep in mind that the statistics presented below refer only to this hard-to-count segment of the population.

Also, the poststrata definition used for PES and Supplement data in this report differ slightly from

that of the official PES estimates. Some whites or homeowners living in predominately black renter blocks may be included in these counts. Black renters residing outside those blocks are excluded.

For more detail on methodology and operations of a PES, read Diffendal (1988).

V. RESULTS

The main results of the Administrative List Supplement project are the tallies for those persons, collectively called the supplement, that can be used in coverage measurement. There are other tabulations that track the progress in identifying the supplement and may be interpreted in trying to understand it. Table 1 shows the results of selecting records in sample blocks from the files supplied by the various agencies.

Operation	Number Kept	%
Procurement	882408	100
In 12 ZIPs	384402	44
Males, 18-44	159235	18
In ALS Blocks	4759	1/2

Table 2 documents the administrative records omitted from further supplement processing. The selection operations were initiated in the order presented in the table, determining, in some cases, which selection culled a record.

Operation	Number Kept	%
Listed for Review	4759	100
Unduplication	2743	56
Sample Addresses	1735	36
Complete Address	1688	35
Missed by PES	1220	26
Unit Occupied	988	21
Precise Age	945	20
Other	916	19

In addition to the types of omission mentioned in the preceding brief overview of the processing, several others worthy of consideration are introduced in the table. A few of the addresses conveyed from administrative sources were missing apartment or unit identification or had a house number between multiunit structures. When the address did not match a single PES address and a followup interviewer would have to enquire at more than three addresses, the record was omitted on the grounds that followup was not feasible or too costly.

Similarly, if the original PES interview determined that an address was unoccupied or if a PES interview could not be conducted there, followup was considered infeasible. That assumption was checked by field visits to a sample of eighteen such addresses. Most addresses were indeed found to be destroyed or otherwise uninhabitable or recently rehhabited or converted to a business. Visits to the other addresses were noninterviews as in the original PES visits. The conclusion of these test visits was that unoccupied status is a reasonable

indication the administrative list person has moved on and that if the PES coded an address noninterview, it likely would remain unproductive.

The selection of cases on the basis of age needed some refinement after poststrata age categories were decided. When records were assigned to poststrata using precise month and year-of-birth cutoffs, very few cases fell outside the two major poststrata. Those few cases would have very little impact on PES estimates and would not represent a full blown effort to supplement those poststrata. They are ignored in subsequent analyses.

A few other omissions are missing the code explaining why they were not processed further. They were discovered after followup was done. Quality control enhancements may ensure such cases are not misdirected or overlooked in future supplement processing.

Table 3 summarizes the result of the sampling of followup cases explained in the preceding section.

Sample Weight	Number
1 (Certainty)	171
4 (Sampled, Selected)	186
0 (Sampled, Not Selected)	559
TOTAL	916

To be used in coverage statistics, supplement persons need to be assigned as matched or nonmatched to census persons. None were to be so assigned unless followup confirmed their status in the census and their belonging in the PES by virtue of residence at a sample address during PES interviewing. Followup did definitely omit some on those criteria. Those omissions are the out-of-scope. Others were unresolved even after the followup interview because it was unclear whether the person might have lived in sample, perhaps at another sample address. Table 4 shows the followup results per followup sampling stratum. Largely because most cases likely to be matched were assigned to the certainty stratum, all final matches were found in that stratum.

Match Status	Followup Sampling Weight		
	1	4	All
Matched	54	00	54
Not matched	37	144	181
Out of scope	38	192	230
Unresolved	42	408	450
TOTAL	171	744	915

The tendency evident in the table for the sampled group to have a higher rate of unresolved interviews lend weak support to using record and address characteristics for identifying the more productive followup cases. Many cases projected to be less productive were resolved and important to the results. Better discrimination functions are needed to separate the productive cases for future practical applications.

The high number of persons unresolved deserves a closer look. If they were resolved, what is the likely effect on the match rate? Are most of the unresolved cases really out of scope and thus not important to the dual system estimation? The breakdown of the types of out-of-scope and unresolved cases in Table 5 helps to answer those questions. One detail not shown in the tables is that few cases changed from likely match to nonmatch or from likely nonmatch to match due to followup and subsequent processing.

Table 5. Breakdown of Unresolved and Out-of-Scope Persons (Unweighted)

<u>Out-of-Scope</u>	
1. Moved out of sample before Census Day	61
2. Moved out of sample between Census & PES	8
3. Died before Census or PES	5
4. Erroneous Census enumeration	12
TOTAL	86
<u>Unresolved</u>	
1. Field followup was not begun	31
2. Knowledgeable respondent was not found	4
3. Respondent unsure if person ever lived there	8
4. Respondent said person doesn't exist	12
5. Respondent unsure when person lived there	7
6. Respondent said person never lived there but was unsure of Census Day address	33
7. Respondent said person once lived there but was unsure of Census Day address	41
8. Respondent said person lived there during Census but was unsure of PES address	8
TOTAL	144

A large proportion of out-of-scope cases is inherent in this data. The source agency may not have undertaken the expensive and difficult task of culling out-of-date records, leaving many movers among those identified at the time of followup as potential supplement cases. The older the record, the more chance for the person to have moved. If agency records were procured long before the census or PES, the records are that much older. Another reason for the high out-of-scope rates is use of a relative's address as a mail drop. Followup interviewer notes show that persons who move around with the military or attending school or following employment or for no stated purpose consider the sample address their permanent address.

The information available for unresolved followup persons is quite varied. At one extreme are those who should have gone to followup but were overlooked. At the other are those persons known to be present at the sample address on Census Day but not certainly during the PES. If the truth were known about unresolved persons, one might expect those about which little was learned to be distributed proportionately over all other levels of resolved and unresolved, controlling for before-followup match status. To distribute those about whom nearly complete information was obtained, a variety of assumptions might be made. Only seven unresolved were likely matches before followup. If the true status of the remaining weighted count of 439 unresolved were fewer than 92% out-of-scope,

the supplement data are conservatively underestimating coverage. It is likely that the rate of out-of-scope persons among the unresolved is higher than the 41% among the resolved cases since uncertainty is likely to be related to absence. Whatever the unresolved cases should be, they either have little effect on the match rates because few of them are in scope or they leave the statistics to be conservative.

Table 6 shows that each list except the Veterans Administration file provided an appreciable number of persons to the supplement. That file just appears to cover a population other than the target group. When considering the cases provided by only an individual source, none of the other lists stands out as either essential or dispensable. In the end it appears to be a case for strength in numbers -- the more lists the better. The best supplement comes from pooling a variety of sources. In addition, the number of times a person is cited on different lists helps identify cases for cost-effective followup sampling.

Table 6. Number of In-Scope and All Followup Records Contributed by Each Source: All or those Only from that Source

Source	In Scope (127 cases)		Followup (357 cases)	
	All	Only	All	Only
IRS	62	12	146	28
Employ. Sec. #1	48	6	122	19
Employ. Sec. #2	46	4	132	27
Motor Veh. Admin.	73	16	196	50
Selective Service	44	6	111	29
Veterans Admin.	0	0	1	0

Having begun with nearly a million administrative records and sifted through them for a relevant handful, the focus changes to what that handful can show. Table 7 presents the resolved in-scope supplement data along with the corresponding PES data and the match rate, the common coverage measurement statistic introduced in the theory section. The table separates the data into the two poststrata covered by the ALS target group: males in black renter blocks between 20 and 29 years of age and those between 30 and 44.

Table 7. PES and Supplement Tabulations and Match Rates by ALS Target Poststratum

	Matched	Total	Match Rate
<u>Males in Black Renter Blocks, Aged 20-29</u>			
Supplement	33	135	24%
Unsupplemented PES	241	323	75%
Supplemented PES	274	458	60%
<u>Males in Black Renter Blocks, Aged 30-44</u>			
Supplement	21	100	21%
Unsupplemented PES	313	384	82%
Supplemented PES	334	484	69%

These match rates are suitable for comparing PES, supplement, and combined tabulations. They do not represent net census coverage. The supplement tabulations are weighted up only by supplement followup sample weights. PES sample design weights which account for block and within-

block subsample size are not used in any of these tabulations. Standard errors for these rates and full dual system estimates are being planned for the complete report on the PES Supplement.

If, as few familiar with the issue seem to doubt, undercount is underestimated in these poststrata, the supplement would appear, using inequality (5) from the theory section above, to improve the coverage estimates of those poststrata. It may be expressed as a comparison of the match rates given in Table 7: if the PES match rate is greater than the separate supplement match rate, the combined match rate is a better coverage measurement statistic than the unsupplemented PES rate.

VI. CONCLUSIONS

A supplement to the Post-Enumeration Survey selected from administrative lists does appear to provide added information about census coverage and census coverage estimation. The 1988 Administrative List Supplement work identified and confirmed one third more PES persons for the target group than the original PES count. That is an accomplishment that other timely extensions of PES procedures have not achieved. And if the undercount is severely underestimated for that target group, response correlation bias should be reduced by supplementing the PES data with sources independent enough to have a higher proportion of persons not found in either PES or census. The 1988 ALS seems to have done so. However, there are some limitations and opportunities to improve this first attempt at an Administrative List Supplement to a Post-Enumeration Survey.

Limitations:

1. Lack of Comparison Data -- There is still no guarantee that the supplemented data are unbiased. It is theoretically possible (although not likely in light of the next point) that the supplemented data overcompensate for response correlation bias and are upwardly biased. Alternate data, as from demographic analysis, do not represent precisely the same population, making comparisons tenuous.

2. Multiple Response Correlation Bias -- The supplement may be subject to the same missing response influences as the census and PES so that responses to all three are correlated to some degree. The field work could be a filter to census, PES, and supplement alike, sifting out all information that tends to be missed in that mode. If so, the supplement merely does not go far enough in removing response correlation bias from the dual system estimate.

3. Followup selections -- It is possible that some address, eliminated from supplement followup because PES called it unoccupied, could have had valid supplement persons.

4. Narrow Focus -- The 1988 ALS was restricted to small segments of the census population. Yet, it did at least provide some boundary information for other groups.

Improvements:

1. Record Vintage -- Extracting records from the source agencies closer to the time of PES interviewing should improve supplement coverage and productivity.

2. Source list selection -- Use every available list which covers the target population.

3. Developing processing -- Efficiency and accuracy may be improved by replacing clerical procedures with computer runs, simplifying tasks, redesigning others to be more like familiar PES tasks, and elaborating quality control procedures.

4. Followup Selections -- The aim is to avoid unproductive, expensive followup visits in the field without passing over valid supplement cases. What proportion of unresolved or out-of-scope results are acceptable? What available variables and discrimination functions can help?

5. Target population -- Expanding the base for the supplement can show relative severity of effects for other groups and improve sex ratio analyses.

6. Use of Demographic Analysis -- If the base for supplement data is as wide as for available demographic data, evaluative comparisons will be more conclusive.

7. Triple System Estimation -- It should extract more information about relationships between census, PES, and ALS (Zaslavsky, 1989).

VII. REFERENCES

- Diffendal, G. (1988), "The 1986 Test of Adjustment Related Operations in Central Los Angeles County," Survey Methodology, 14, 71-86.
- Fay, R.E., Passel, J.S., and Robinson, J.G. with Cowan, C.D. (1988), "The Coverage of the Population in the 1980 Census," 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Bureau of the Census, Washington D.C.
- Marascuilo, L.A. and McSweeney, M. (1977), Nonparametric and Distribution-Free Methods for the Social Sciences, Monterey, CA: Brooks/Cole.
- Marks, E.S., Seltzer, W., and Krotki, K.J. (1974), Population Growth Estimation, New York: The Population Council.
- Wolter, K.M. (1986), "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-346.
- Zaslavsky, A. (1989), "Multiple-System Methods for Census Coverage Evaluation," Proceedings of the Fifth Annual Research Conference, Bureau of the Census, Washington, D.C.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

** All records received from outside the Census Bureau are handled with care for the confidentiality and other specified requirements of all agencies involved.