UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

August 9, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-4

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: David Whitford and Rita Petroni
Chairs, A.C.E. Revision II Measurement Subgroup

Prepared by: Douglas Olson,
Mathematical Statistician,
Decennial Statistical Studies Division

Subject: A.C.E. Revision II – Contingency Sample Design for Clerical Recoding

One operation of the Accuracy and Coverage Evaluation (A.C.E.) Revision II project is to correct measurement error for a subsample of the A.C.E. We will attempt to correct for measurement error using the best available information, whether from the A.C.E. Person Follow-Up (PFU) or the Evaluation Follow-Up (EFU). The goal is to obtain better codes for all of the sample for which an evaluation follow-up was conducted. For a large portion of the data, the recoding can be automated. For the rest of the data, the recoding will need to be done by our "best coders" or analysts at the National Processing Center (NPC). We have been allotted twelve weeks of the analysts' time for this A.C.E. Revision II recoding operation. Given prior experience, we anticipate that about 24,000 cases can be handled in this time frame. In the event that the entire workload cannot be completed, a contingency sampling plan was developed and embedded in the analysts' workflow. If the entire workload is finished, this sampling plan will not be used and if it is not finished, the sampling plan will be triggered. This memorandum documents the contingency sampling plan and identifies related estimation issues.

Background

The evaluation follow-up was conducted for the entire Measurement Error Reinterview (MER) sample. This sample consists of about 70,000 persons from both the P and E samples, drawn from 2,259 block clusters. The clusters were systematically sampled at differential rates among

the 11,303 clusters in the A.C.E. The persons who were included in the A.C.E. and who lived in the selected clusters were included in the MER sample except:

- Some types of persons not eligible for follow-up were excluded from the MER
- Persons in households that were wholly matched were subsampled at 1-in-7
- Persons in households that were wholly non-matched were subsampled at 1-in-5

Every person in the MER clusters was weighted by the inverse of the probability of selection. Generally speaking, the MER weight of any cluster member was the product of their weight in the A.C.E. divided by the probability with which their cluster was selected into MER. Subsampled persons were assigned an additional weight of 7 or 5 if they were included in the sample, 0 if excluded. Persons ineligible for follow-up, and hence excluded from MER, were also zero-weighted. It was possible for a person's TES status to change as a result of the EFU interview, so a factor of 4.8925, the TES sampling weight, could be either applied to or removed from the MER weight of persons in TES sampled clusters.

Sampling Plan Overview

At the time this sampling plan was developed, the sample included 23,988 clerical cases. Subject matter experts advised that we must plan for the contingency that only 80 to 90 percent of this workload might be completed, where workload is proportional to the number of clerical cases. Of the 2,259 total clusters, 166 were dropped because they contained no people; another 246 have no clerical cases, leaving 1,847 with clerical cases to constitute the sampling universe. The clusters represent an approximately random cross-section of the A.C.E., excluding certain persons ineligible for the Evaluation Follow-Up interview.

Combining the sample universe size (1847 clusters), daily workload capacity (50 clusters) and expected total workload capacity (80 to 90 percent of certainty cases) suggested a targeted sample design with:

- A total of 847 certainty clusters containing approximately 50 percent of clerical cases
- A total of 1,000 non-certainty clusters, randomly assigned into twenty batches of 50 clusters each

Among the characteristics desired in the final sample:

- Contain most of the total weight for the MER sample
- Contain an adequate sample size of all race domains, and of inmovers and outmovers
- Contain the clusters with the most Clerical cases

To this end, the following certainty criteria were applied in order:

- The 10 clusters with the most race Domain 1 (American Indians on Reservations) total weight
- The 10 clusters with the most race Domain 5 (Native Hawaiians and Pacific Islanders) total weight
- Clusters with at least 50 Clerical cases
- Clusters with at least one person weight of 50,000 in either the E or P sample
- The clusters with the highest total weight using the definition below to get a total of 847 certainty clusters

The term "total weight" refers to the combined weighted total number of E and P-sample person records based on the MER final weight.

After the 847 certainty clusters were selected, the remaining 1,000 clusters were sorted in the same order in which they were sorted for selection into the Evaluation Follow-up sample. The fields STRATUM (a poststratum that partitions the A.C.E. universe of clusters) and N (an ordinal listing of housing units within STRATUM), copied from the BFUSAMP file maintained by PRED, sort the records into the desired order. They were assigned into batch numbers 1 to 20, recycling after every 20th record, into the 20 pre-randomized batches.

Rationale for Certainty Rules

To insure adequate sample size for all race domains, it was decided that race Domains 1 (American Indians on Reservations) and 5 (Native Hawaiians and Pacific Islanders), needed their own certainty selection because they are the smallest domains, are heavily clustered and have relatively low total weight. Analysis after drawing a test set of certainty requirements showed that Domain 2 (American Indians off Reservations), which is the next-smallest domain, did not require special certainty selection to insure a representative sample size, nor did any larger domains.

The rule about 50 Clerical cases was developed because it was initially suggested to generate certainty clusters that included 25 percent of the total clerical case workload based on the top workload-count clusters. Such a criterion would have selected 68 clusters with at least 51 cases totalling 5,997 total cases. Expanding the definition to 50 cases changed these totals very little while making the criterion easier to define.

The maximum weight criteria (50,000) was set so that if the sampling rate was not smaller than 1-in-2, no new weights greater than 100,000 could be generated by sampling. Since we expect to complete at least 80 percent of the workload and about 55 percent is represented by the certainty cases, it should be possible to complete at least half of the noncertainty cases and assign them a weight no more than 2.

The certainty clusters have been assigned a random ordering (1 to 847) because of concerns by subject matter experts that coders could experience a learning curve over the course of the project, which could introduce non-sampling error if some types of clusters were performed before others.

Using the Sample

The batch sampling plan assumes that all certainty cases will have their recoding completed and then as many sample batches will be completed as time allows. It is possible, as of this writing, that all 20 batches will be completed and that this sampling plan would not be used. If this does not happen, and the sampling plan does need to be implemented, sampling weights will be calculated.

Weighting Issues

As of this writing, one fundamental issue about weighting the batch sample has not been determined. If some clusters do not have their clerical operations performed and the sample is implemented, either:

1) Clusters in which clerical work is not performed would be dropped from the sample in their entirety and represented by application of sampling weights to the completed sample clusters.

2) Clusters in which clerical work was not performed would still have their non-Clerical cases used. The clerical cases in clusters performed would be weighted to represent the clerical cases of the not-performed clusters.

In either case, all persons in Certainty clusters would be weighted 1.00.

The sampling fraction from the Batch sampling would be batches completed divided by 20. The factor would be applied to the weights from the other phases of sampling.

Variance Estimation

Variance estimation methods will reflect the additional weighting resulting from the Batch sampling. It will not reflect the covariances between sample clusters, since each cluster effects the selection probability of the other clusters by only 1/1000, an effect too small to be worth the effort of writing into variance estimation programs.

Characteristics of the Batch Sample

Table I (see Appendix) shows important characteristics of the certainty cases and the PRE–RANDOMIZED sampled batches. The certainty clusters included:

55 percent of Clerical cases
56 percent of non-Clerical cases
48 percent of the American Indians on Reservations E-sample
57 percent of the American Indians on Reservations P-sample
78 percent of the Native Hawaiian and Pacific Islanders E-sample
73 percent of the Native Hawaiian and Pacific Islanders P-sample
63 to 87 percent of the remaining domains: American Indians off Reservations 63 percent, Hispanics 76 percent, Asians 63 percent, Blacks 83 percent and White and Other Race 87 percent E-samples
83 percent of the total sample weight for both the P- and E-samples

Not shown in the tables is that 77 percent of both inmovers and outmovers are included with certainty, Inmovers (10.332 of 13.416 million), Outmovers (7.226 of 9.356 million).

The sampling plan distributes the workload quite well, with each of the twenty batches containing between 451 and 604 clerical cases (4.1 percent to 5.6 percent of all clerical cases in the sample.)

The sample batches will be performed in order: 10, 12, 2, 1, 13, 15, 19, 5, 16, 6, 9, 20, 8, 17, 3, 14, 18, 11 , 7, 4. The certainty clusters have also had a random order assigned to reduce the likelihood of non-sampling error caused by coders experiencing a learning curve during the earlier parts of the operation.

File Layouts

Batsamp: (SAS Dataset containing only minimal information for ordering)
Delivered to staff members to communicate to the National Processing Center for operational use
2,093 Records; one per cluster requiring work; sorted on Certord x Packord x Clust

Field	Type	Description
Clust	Char 6	6-digit cluster number
Packord	Numeric	Randomized batch order
Certord	Numeric	Randomized certainty cluster order (=0 if no clerical cases)

BatchSDF: (SAS Dataset containing variables and data used in generating the sample)
File maintained by staff who created the sample as reference for sampling methodology
2,093 Records; one per cluster requiring work; sorted on Clust; layout in attachment.

Table I
A.C.E. Revision Batching Sample Design

Clusters	Clerical Cases	Non-Cler Cases	Domain 1 E-sample	Domain 1 P-sample	Domain 5 E-sample	Domain 5 P-sample	Domain 2 E-sample	Domain 3 E-sample	Domain 4 E-sample	Domain 6 E-sample	Domain 7 E-sample	P-sample	E-sample
Certainty Batches													
No Clerical cases													
246	0	1,841	17,651	34,119	0	4,975	13,754	1,014,369	934,676	775,607	15,494,747	18,883,429	18,250,805
Top 10 Domain 1													
10	243	775	180,590	248,789	0	0	0	12,991	0	0	6,794	270,330	200,374
Top 10 Domain 5													
10	140	394	0	0	204,726	244,377	1,981	430,445	7,222	239,797	467,188	1,510,218	1,351,360
50+ Clerical cases													
74	6,159	10,871	0	0	9,263	24,912	31,140	2,740,472	2,460,649	387,334	5,021,240	10,802,043	10,650,099
Top Total Weight or single weight over 50,000													
753	6,570	24,566	0	0	75,046	81,695	824,369	20,389,559	15,665,369	6,690,462	148,597,444	206,908,959	192,242,248
Sample Universe													
1,000	10,876	30,150	212,725	214,525	82,551	134,505	518,402	7,731,934	11,330,779	1,604,248	24,789,325	48,818,222	46,269,965
Sample Batches													
1	50	601	1,511	9,178	4,005	22,749	24,060	58,961	254,699	580,477	68,789	1,199,770	2,228,178	2,194,624
2	50	598	1,728	5,952	4,433	10,502	12,076	21,830	592,709	474,206	127,300	1,262,172	2,398,902	2,494,670
3	50	501	1,308	5,254	4,951	1,354	1,230	54,010	242,990	738,483	33,230	1,311,713	2,613,498	2,387,035
4	50	560	1,388	6,743	8,649	4,433	7,397	13,610	406,359	696,847	77,261	1,167,127	2,665,337	2,372,380
5	50	586	1,270	26,598	24,792	0	0	2,292	337,710	485,983	105,926	1,001,548	2,303,654	1,960,057
6	50	542	1,442	8,545	13,116	1,005	11,329	17,216	451,042	705,330	187,617	1,259,469	2,714,303	2,630,224
7	50	584	1,645	30,586	31,598	1,640	2,514	78,912	243,987	472,051	75,021	862,477	2,290,545	1,764,673
8	50	469	1,543	6,358	8,577	12,087	9,700	26,986	334,989	624,783	65,188	1,233,767	2,200,033	2,304,158
9	50	522	1,597	31,156	25,047	6,068	23,866	11,069	429,200	560,477	57,743	1,514,022	2,458,587	2,609,734
10	50	551	1,589	7,266	8,955	3,876	4,191	17,509	563,296	531,344	62,478	1,122,170	2,551,467	2,307,940
11	50	596	1,319	0	0	9,602	9,042	21,465	401,139	538,778	49,938	1,464,637	2,877,769	2,485,560
12	50	451	1,334	6,643	6,469	3,518	7,341	13,113	382,851	368,419	76,463	1,290,049	2,317,069	2,141,056
13	50	573	1,418	2,933	4,306	0	0	36,925	353,902	619,772	85,893	1,248,225	2,352,149	2,347,649
14	50	536	1,494	13,128	14,159	3,018	4,053	1,328	228,009	506,731	70,035	1,418,188	2,395,137	2,240,438
15	50	604	1,818	15,635	17,459	1,257	6,943	9,012	257,584	542,081	82,682	1,421,639	2,167,265	2,329,890
16	50	566	1,630	7,871	8,431	714	8,543	29,068	186,145	584,451	124,574	1,383,639	2,462,448	2,316,463
17	50	531	1,805	0	0	37	1,039	57,394	588,432	349,545	16,191	1,202,204	2,347,952	2,213,803
18	50	576	1,712	8,830	11,201	644	0	41,286	691,801	484,288	45,726	1,181,790	2,526,189	2,454,366
19	50	465	1,500	14,147	12,943	0	677	4,754	490,043	770,456	155,129	1,174,569	2,730,857	2,609,098
20	50	464	1,099	5,902	5,436	48	503	1,662	295,045	696,276	37,062	1,070,152	2,216,881	2,106,147

Attachment

Layout of SAS Dataset BatchSDF, used in developing Batch Sampling

#	Variable	Type	Len	Pos	
1	CLUST	Char	6	336	
From Both2 (file created by staff summarizing relevant data values):					
2	stratum	Char	7	342	EFU Sampling Stratum
3	clericalw	Num	8	0	Weighted Clerical Cases
4	notclerw	Num	8	8	Weighted Non-Clerical Cases
5	clericaluw	Num	8	16	Unweighted Clerical Cases
6	notcleruw	Num	8	24	Unwgtd Non-Clerical Cases
7	nmcount	Num	8	32	Weighted Non-matches
8	nrcount	Num	8	40	Weighted Non-residents
9	inmcount	Num	8	48	Weighted Inmovers
10	eecount	Num	8	56	Weighted EE's
11	ecount	Num	8	64	Weighted E-sample
12	domain1	Num	8	72	Weighted P&E Domain Totals
13	domain2	Num	8	80	"
14	domain3	Num	8	88	"
15	domain4	Num	8	96	"
16	domain5	Num	8	104	"
17	domain6	Num	8	112	"
18	domain7	Num	8	120	"
From EMDVF and PMDVF (EFU Results files):					
19	edom1	Num	8	128	Weighted E-sample Domain Totals
20	edom2	Num	8	136	"
21	edom3	Num	8	144	"
22	edom4	Num	8	152	"
23	edom5	Num	8	160	"
24	edom6	Num	8	168	"
25	edom7	Num	8	176	"
26	maxewgt	Num	8	184	Maximum MER E-weight
27	pdom1	Num	8	192	Weighted P-sample Domain Totals
28	pdom2	Num	8	200	"
29	pdom3	Num	8	208	"
30	pdom4	Num	8	216	"
31	pdom5	Num	8	224	"
32	pdom6	Num	8	232	"
33	pdom7	Num	8	240	"
34	maxpwgt	Num	8	248	Maximum MER P-weight
35	etot	Num	8	256	Total weighted E-sample
36	ptot	Num	8	264	Total weighted P-sample
37	petot	Num	8	272	Total wght E&P samples
Sampling Program Generated:					
38	certgrp	Char	1	349	Certainty Reason A: Zero Clerical cases B: Top 10 Domain 1 C: Top 10 Domain 5 D: Maximum weight>50,000 E: Top total Weight Z: Sample Case
39	samgrp	Num	8	280	Pre-randomized packet #1-20
40	packord	Num	8	288	Randomized packet Order 1-20
41	certord	Num	8	296	Order to perform Certainty 1-843
42	rannum	Num	8	304	Random Number used in Sorting
From BFUSAMP (MER sample selection file):					
43	digit	Char	1	350	Cluster check (sixth) digit
44	n	Num	8	312	Within-stratum order from original EFU sampling
From PMDVF (P-sample MER file):					
45	inmov	Num	8	320	Weighted in-movers
46	outmov	Num	8	328	Weighted out-movers

Note: All P-sample weights are sums of non-movers, inmovers and outmovers without regard to residence probability