AN EMPIRICAL INVESTIGATION OF ALTERNATE ITEM NONRESPONSE ADJUSTMENTS*

Brenda G. Cox and Ralph E. Folsom
Research Triangle Institute

In this empirical investigation of alternate item nonresponse adjustments, two methods which are frequently used by statisticians to adjust for bias induced by item nonresponse were studied. In particular, the hot deck and weighting class adjustment techniques were compared using data from the National Longitudinal Survey of the High School Class of 1972. Estimates obtained using these two techniques were compared with respect to their bias, variance, and mean square error to estimates obtained when no item nonresponse adjustments were made.

1. Construction of the Experimental Data Set

Rather than constructing a data set with artificially induced nonresponse, the decision was made to use actual data that contained item nonresponse for which the answers were subsequently obtained by telephone followup activities. By using data with naturally occurring patterns of item nonresponse, it was felt that a better understanding could be obtained of the actual problems associated with item nonresponse and the effect of nonresponse adjustments on the precision of the resulting estimators. Such a data set was constructed from the NLS Third Followup (TFU) Survey 1/ by taking account of the following set of special circumstances. Certain items on the questionnaire were designated key items by NLS staff. When an incoming questionnaire had a missing response or an inconsistent set of responses for one or more of these key items, the questionnaire was marked as having failed edit, the individual involved was telephoned, and the missing response(s) added from the telephone interview. The data records for individuals whose questionnaire failed edit contained the responses to these critical items but did not indicate which responses were obtained by telephone, or what the original responses were.

In order to obtain this information on the responses before telephone resolution, the questionnaires were re-examined by data editors and the original responses to the selected key items recorded. In all, a total of 10,850 questionnaires failed edit. For reasons of economy, a subsample of size 5,854 was selected for re-examination. Twenty key items chosen to be representative of the NLS instrument were examined on each of the selected questionnaires and a notation made as to whether or not telephoning was necessary to obtain a response to that particular item. They include questions that have categorical responses including four items which allow the student to choose multiple response options. Many should have been answered by all survey participants; others applied to subpopulations such as those employed or in school. Some questions were included that came from skip or "within-routing" patterns. Other sensitive questions, such as income, were included which had quantitative responses.

For each of these twenty items, the status of the response before telephone followup was determined. A summary of the status of the original responses to these items for the sample of"fail-edit"questionnaires is given in Table 1. Except for two multiple response option questions (TQ1 and TQ9) and four financial questions (TQ89A, TQ89B, TQ141A, and TQ141B), over ninety percent of the questionnaires contained a response for an item that was consistent with other responses on the questionnaire. The highest rates of missing or blank responses were found for the income items TQ141A and TQ141B. The multiple response items, TQ1 and TQ9, had the highest inconsistency rates; that is, TQ1 and TQ9 responses were most frequently in conflict with other questionnaire items. The "other" category in Table 1 is composed of those who failed an item but could not be contacted for telephone resolution.

The results presented in Table 1 are based upon the subsample of size 5,854 drawn from the 10,850 questionnaires that failed edit. Adding in the 9,235 questionnaires which passed edit (and hence had "consistent" answers for all of these items) would reduce all of these percentages by about half. Thus the data set that was constructed of original responses had a relatively small rate of item nonresponse and a somewhat larger rate of inconsistent responses. Judging from where the inconsistencies occurred, the major problem,other than TQ1 and TQ9,appeared to be associated with the"routing pattern"questions. Thus,in reconstructing the data set, the decision was made to leave the inconsistent data as observed rather than to code inconsistent items as blank. The hot deck program and the weighting class imputation program were then written to force consistency on the data by requiring that the responses within a routing pattern agree with the lead-in question to the routing pattern. In computing the no-imputation estimates, no attempt was made to force consistency on the data within records.

2. The Hot Deck and Weighting Class Imputation Techniques

The hot deck technique is flexible and relatively inexpensive to run with respect to computer time. Before using the hot deck imputation procedure, the data file was sorted into 87 weighting classes and then according to strata and school within strata. The weighting classes which were based upon the student's race, sex, high school grades, high school curriculum, and parents' education,were originally formed for total questionnaire nonresponse adjustments. These weighting classes were adapted for item nonresponse imputation by incorporating certain routing pattern lead-in questions. For each weighting class, an initial hot deck was formed by going through the data file and recording the first completed response to each item. Then, as the new data was processed, the weighting class to which each individual belonged was determined. If the item examined was complete, then that individual's response replaced the response stored in the hot deck for that weighting class. Thus new responses were supplied for the hot deck as they appeared in the data file. When a questionnaire with a missing item was encountered, the response in the hot deck for that weighting class was imputed for the missing response. 2/

The second item nonresponse adjustment technique used was a"weighting class"imputation method. Only the technique as applied to continuous variables will be discussed in this paper. For continuous variables, the weighting class imputation technique simply replaced missing values by the estimated respondent mean for the weighting class containing the individual. 3/

3. Analysis of the Data Set

Estimates of means and proportions were obtained using both imputation procedures for the whole population and domains defined by race, sex, ability, socioeconomic status, region, and race-by-ability. Cross-tabulations were examined to determine the effect of item nonresponse imputation methods on multivariate statistics.

The variance of the sample means and proportions were estimated using the balanced repeated replication technique (BRR). BRR uses a balanced set of half-sample estimates to compute the sampling variance of complex statistics. The variability among the replicated estimates approximates the desired variance (McCarthy, 1966). In this investigation, sixteen equal-sized superstrata were formed and the item nonresponse imputation procedures were applied separately to the associated set of sixteen balanced half-samples; this insured that the resulting BRR variance estimates reflect the variability induced by the imputation procedures.4/

4. Summary of Initial Results

Due to the high item response rates, statistics computed from the pre-telephone data set had a relatively small amount of bias when compared with estimates using the post-telephone followup data, corrected and completed. The bias that was observed resulted from two response error sources; namely, nonresponse or missing items, and inconsistent responses. In this investigation, no general attempt was made to force consistency on the data within a student's record. An exception was made in the hot deck and weighting class imputation programs which did force the responses to items within a routing pattern to agree with the lead-in question to the routing pattern.

Table 2.--In general, the hot deck procedure did appear to reduce, for discrete items, the bias caused by nonresponse. The greatest improvement in bias was seen with respect to item TQ118 which also exhibited the most nonresponse of the discrete items. Results for the proportion of students responding "3" to TQ118 are given in Table 2. The table gives the sample size for each domain, the "true" value of the statistic estimated using the telephone corrected and completed data file, the relative bias(RB) of the hot deck (HD) and no-imputation (NI) procedures, and the root mean square errors of the procedures. Note that most, if not all, of the gain in bias reduction from hot decking was lost by a corresponding increase in the variance of the estimates.

Table 3.--The hot deck technique does not appear to perform very well for continuous items including the income questions which had the highest rate of item nonresponse. In general, the hot deck imputations did not improve estimates of means, and again, a compensating increase in the variance of the hot deck estimates tended to counteract the bias reduction that was occasionally obtained.

Since the weighting class estimates for the discrete items have not yet been completed, only the results for items with continuous responses will be discussed. Overall, the weighting class imputation technique performed best for the income items, TQ141A and TQ141B, which exhibited the most nonresponse. The weighting class estimates had somewhat smaller mean square errors than the no-imputation and hot deck procedures.

Table 4.--Due to the manner in which the data file was constructed, it was relatively easy to identify inconsistent items. Recognizing that measurement errors caused by inconsistent responses constitute an important source of bias in the estimates obtained using the pre-telephone file, a new data file was constructed which retained all the inconsistent responses from the mail questionnaire but which had missing items replaced by responses obtained in the telephone followup. The difference between statistics using this missing-data-corrected file (referred to as $\bar{Y}_{ME}$ where ME stands for measurement error) and statistics obtained from the fully corrected telephone followup data file (referred to as $\bar{Y}_{TRUE}$) provides an estimate of the measurement error associated with inconsistent responses. Referring to Table 4, which compares these two statistics, one can see that the measurement error associated with inconsistent responses had a significant effect for TQ1 and TQ9. TQ1 and TQ9 were multiple response option questions in which the students were instructed to "Circle as many as apply to you." The measurement error associated with these results was large and positive indicating that many students failed to circle all the options that applied to them. Note that on the far right in Table 4 are the estimates $\bar{Y}_{NIC}$. These were obtained using no imputation on the pre-telephoning data set where inconsistent responses were recoded as missing data. For the questions in which inconsistencies were most common, i.e., TQ1 and TQ9, $\bar{Y}_{NIC}$ produced less biased estimates than $\bar{Y}_{NI}$ (no imputation) and $\bar{Y}_{HD}$ (hot deck) obtained using the pretelephone data set with the inconsistencies left in.

5. Conclusions

In general, no significant gains in precision were achieved by using the imputation procedures. In part, this was because the response rates for the individual items were quite high. Also, the lack of important gains through imputation can be attributed to the fact that a reduction in bias was accompanied by a compensating increase in variance. If the response rate had been smaller and the associated nonresponse bias larger, the effect of bias reduction might have more than offset the corresponding increase in the variance of the statistics. For the continuous items where weighting class estimates could be compared with no imputation and hot decking, the weighting class estimates did have somewhat smaller mean square errors for the items with higher nonresponse rates. Unfortunately, accurate variance approximations for imputation-based statistics are difficult and costly to obtain so that most users will ignore the imputation in computing the variance. In a sense then, one disadvantage of using imputation techniques will be to underestimate the true variance of sample statistics (Ford, 1976; and Bailar and Bailar, 1978). This underestimation could jeopardize the validity of confidence statements.

Finally, it should be emphasized that this study focused exclusively on the estimation of univariate means and proportions. If more complex statistical analyses were being conducted, such as regression or factor analysis using many variables, it might be much easier to analyze the data when the missing values have been imputed. Also, the effect of item nonresponse is probably cumulative so that even though individual items have a large response rate, the number of records with complete responses to all the items entering into the analysis may be so small that some type of imputation becomes necessary. The effects of imputation on inference when more complex statistical analyses are performed is a topic deserving considerable further investigation.

6. Footnotes

* This research was supported by the National Center for Education Statistics under contract number OEC-0-73-6666.

1/ A general description of the National Longitudinal Survey and a copy of the TFU survey instrument may be found in Levinsohn (1978). The results presented in this paper are preliminary findings from an NLS methodological study. The final report for this study will be available from the National Center for Education Statistics at a later date.

2/ Note that since the data file was sorted into weighting classes before imputing for missing values, one would expect the hot deck technique to obtain much if not all of the bias reduction that would have resulted from the weighting class adjustment procedure, but with a somewhat larger variance.

3/ When a mean or total is being estimated, this weighting class imputation technique results in the same estimate as that obtained when weight adjustments are made within weighting classes. A more complete description of these two techniques and other nonresponse imputation and adjustment techniques may be found in Chapman (1976) and Bailar (1977).

4/ While no formal justification has been developed for using BRR in this context, we feel that such half-sample estimates should reflect the sampling variability of the imputation-based statistics.

7. References

[1] Bailar, B. A., Bailey, L., and Corby, C. (1977). A comparison of some adjustment and weighting procedures for survey data. Paper presented at University of North Carolina Symposium of Survey Sampling and Measurement, April 14-17, 1977.

[2] Bailar, John C. and Bailar, Barbara A. (1978). Comparison of Two Procedures for Imputing Missing Survey Values. American Statistical Association Proceedings, Section on Survey Research Methods, 1978.

[3] Chapman, David W. (1976). A survey of non-response imputation procedures. American Statistical Association Proceedings, Social Statistics Section, 1976, pp. 245-251.

[4] Ford, Barry L. (1976). Missing data procedures: a comparative study. American Statistical Association Proceedings, Social Statistics Section, 1976, pp. 324-329.

[5] Levinsohn, J., Lewis, L., Riccobono, J. A., and Moore, R.P. (1978). National Longitudinal Survey: Base year, first, second, and third follow-up data file users manual. National Center for Education Statistics, DHEW.

[6] McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys. National Center for Health Statistics, Series 2, No. 14.

Table 1. Classification of Original Responses for Fail Edit Questionnaires (in percent)

Item	Originally Consistent	Resolved by Telephone		Other*
		Originally Blank	Originally Inconsistent	
Discrete				
TQ1	86.0	0.3	12.4	1.4
TQ9	76.1	0.5	21.6	1.8
TQ10	92.6	1.6	4.9	0.8
TQ12	95.0	0.9	3.6	0.5
TQ29	98.0	1.0	0.6	0.4
TQ33	91.0	0.9	7.4	0.8
TQ51	95.0	1.7	2.2	0.7
TQ52	93.0	1.6	4.7	0.8
TQ66	90.0	1.9	7.3	1.0
TQ90	95.1	2.6	1.6	0.7
TQ101	97.3	1.8	0.4	0.5
TQ102	98.3	0.4	0.9	0.3
TQ118	94.7	4.4	0.2	0.8
TQ129	98.6	1.0	0.2	0.3
TQ131	99.1	0.4	0.3	0.2
TQ136	99.2	0.4	0.2	0.2
Continuous				
TQ15	97.5	1.1	0.9	0.5
TQ16	94.9	3.3	1.1	0.7
TQ89A	86.8	4.8	7.0	1.3
TQ89B	86.0	5.1	7.4	1.4
TQ141A	68.2	23.8	4.4	3.5
TQ141B	67.6	23.9	4.9	3.6

*Unresolved cases, i.e , originally inconsistent or blank responses not resolved by telephone.

Table 2. A Comparison of Hot Deck vs. No Imputation for TQ118, Response No. 3

Subpopulation		$\bar{Y}_{TRUE}$	NI RB%	HD RB%	NI $\sqrt{MSE}$	HD $\sqrt{MSE}$
Total		6.42	1.09	0.16	0.26	0.25
Sex:	Male	11.44	0.35	0.17	0.46	0.46
	Female	1.24	0.81	0.00	0.15	0.15
Race:	Black	9.72	3.09	-1.34	0.96	0.96
	White	5.90	1.02	0.51	0.26	0.25
	L. Am.	8.61	1.51	13.01	5.02	5.12
Ability:	Low	7.05	2.13	-0.43	0.59	0.53
	Middle	6.27	0.48	-0.16	0.32	0.35
	High	5.17	1.55	1.74	0.39	0.40
SES:	Low	8.33	0.96	0.36	0.61	0.62
	Middle	6.46	1.55	-0.16	0.31	0.30
	High	4.03	1.49	0.74	0.45	0.44
Region:	NE	5.36	1.31	0.37	0.92	0.90
	NC	5.96	0.67	-0.67	0.47	0.52
	S	7.17	0.98	0.28	0.28	0.30
	W	7.53	1.86	1.06	0.28	0.28
Race x Ability:						
Black:	Low	7.58	5.15	1.58	0.95	0.86
	Middle	12.21	-2.54	-5.49	1.81	2.45
	High	9.63	3.01	0.00	6.43	6.34
White:	Low	6.67	1.50	-0.75	0.55	0.56
	Middle	5.90	0.85	0.45	0.40	0.41
	High	5.10	1.57	1.76	0.37	0.38

The first column shows the corrected or completed responses, i.e., Y_{TRUE}. The relative bias (RB) is shown for no imputation (NI) and hot deck (HD) procedures in the second and third columns. The last two columns show the root mean square error, $\sqrt{MSE}$, of the no imputation and hot deck methods.

Table 3. A Comparison of Hot Deck vs. No Imputation for TQ141A

Subpopulation		$\bar{Y}_{TRUE}$	NI RB%	HD RB%	WC RB%	NI $\sqrt{MSE}$	HD $\sqrt{MSE}$	WC $\sqrt{MSE}$
Total		7040	-0.44	0.66	-0.16	76.25	96.54	70.99
Sex:	Male	6623	0.02	0.70	0.09	92.32	106.12	95.38
	Female	7460	-0.64	0.61	-0.39	146.36	178.25	141.99
Race:	Black	5946	-3.03	-0.62	-0.90	250.67	187.38	165.74
	White	7139	-0.54	0.67	-0.16	72.54	95.38	65.43
	L. Am.	8927	-4.01	-6.09	-8.72	1195.68	1108.22	1251.94
Ability:	Low	7993	0.30	2.34	-0.53	165.32	291.28	156.29
	Middle	7574	0.75	0.28	0.25	100.94	95.61	79.44
	High	5327	-0.14	1.08	0.38	111.18	129.07	108.87
SES:	Low	7663	-0.71	1.19	-0.64	136.33	190.76	131.91
	Middle	7585	-0.47	0.49	-0.36	89.73	100.16	86.08
	High	5346	0.52	0.46	1.05	112.84	155.40	123.44
Region:	NE	6542	-1.51	-1.20	-0.62	189.22	195.43	163.25
	NC	7353	-0.80	0.81	-0.62	102.06	132.73	95.81
	S	7109	0.57	2.20	0.60	161.28	262.98	151.61
	W	7137	-0.05	0.35	0.00	132.00	149.82	118.44
Race x Ability:								
Black:	Low	6438	-4.11	-0.86	-2.44	320.17	253.81	210.98
	Middle	5648	3.26	1.78	3.62	446.45	429.52	433.24
	High	2884	12.56	6.89	34.38	740.93	570.20	1241.35
White:	Low	8578	-0.81	2.23	-1.21	171.07	313.05	178.87
	Middle	7687	0.65	0.16	0.17	91.03	97.86	70.24
	High	5380	-0.30	1.06	0.17	101.50	101.30	98.62

This table is labelled in a manner similar to that for table 2 except that the relative bias (RB) and root mean square error ($\sqrt{MSE}$) are shown here not only for the no imputation (NI) and hot deck (HD) methods but also for the weighting class (WC) procedure.

Table 4. A Comparison of Various Estimators of Proportions for the Total Population

Item	Response	$\bar{Y}_{TRUE}$	$\bar{Y}_{ME}$	$\bar{Y}_{NI}$	$\bar{Y}_{HD}$	$\bar{Y}_{NIC}$
TQ1	1	72.29	70.55	70.56	70.55	72.98
TQ9	1	67.82	63.33	63.28	63.29	67.15
TQ10	1	61.22	61.28	61.28	61.24	62.45
TQ10	2	13.06	13.01	13.03	13.06	11.81
TQ10	3	1.45	1.39	1.37	1.35	1.24
TQ10	4	24.27	24.32	24.33	24.35	24.51
TQ12	1	25.66	27.04	26.95	26.50	22.38
TQ12	2	7.20	7.19	7.24	7.23	7.59
TQ12	3	67.14	65.77	65.81	66.27	70.03
TQ118	1	92.68	92.68	92.59	92.67	92.63
TQ118	2	0.90	0.91	0.92	0.90	0.88
TQ118	3	6.42	6.41	6.50	6.43	6.49
TQ129	1	9.78	9.80	9.77	9.80	9.73
TQ129	2	45.95	45.95	45.92	45.92	45.93
TQ129	3	4.02	4.01	4.02	4.01	4.02
TQ129	4	40.25	40.24	40.30	40.28	40.32

The first column shows the estimate $\bar{Y}_{TRUE}$ obtained from the fully corrected and completed telephone file. $\bar{Y}_{ME}$ stands for the data after correcting for nonresponse but not for inconsistencies. $\bar{Y}_{NI}$ is the no imputation data with inconsistencies left in. $\bar{Y}_{HD}$ is the hot deck estimate obtained using the pre-telephone data set with the inconsistencies left in. $\bar{Y}_{NIC}$ is the no imputation estimate with the inconsistencies coded as missing.