An Analysis of Missing Income Information on the General Social Surveys

Tom W. Smith

NORC

University of Chicago

GSS Methodological Report No. 71

October, 1991

This research was done for the General Social Survey Project directed by James A. Davis and Tom W. Smith. The project is funded by the National Science Foundation, Grant No. SES-87-18467. Item non-response creates two problems for survey analysts. First, the missing data reduces the sample size. This increases sampling variance and may hinder certain analyses such as subgroup comparisons. Second, item nonresponse may create non-response bias. If the values of the missing data differ from that of the known data, then it biases the known data and makes it unrepresentative [Endnote 1]. Assessing item non-response bias is usually difficult, since the values of the missing cases are unknown.

Missing Income Data

For most demographics on the General Social Surveys (GSS) [Endnote 2], item non-response is a trivial problem since values are ascertained for nearly all cases. As Table 1 shows, for 15 of 18 standard demographics non-response is below l%. The exception to this pattern is household income which has a non-response rate of 8.2%. There are more missing data for household income than for all of the other standard demographics combined and the two variables next highest on the list, relative income level of respondent's family of origin and numbers of household members earning money last year, are both income-related variables.

In addition, the amount of missing income data on the GSS has significantly increased over time. Non-response averaged 7.1% in 1973-75, 8.5% in 1980-1982, and 10.1% in 1989-91.

The missing income problem is not restricted to the GSS, but is common to virtually all surveys. In fact, the level of missing income data on the GSS is typically lower than on many other surveys both in the United States (Table2A) and in other countries (Table 2B).

Reasons for Missing Income Data

The difficulty of obtaining income data centers around two problems. First, there is a reluctance to share income information with others (e.g. with interviewers). Income information is typically considered to be private and personal (Bradburn and Sudman, 1979). Some people are opposed to sharing the information, simply because they define it as "nobody's business". In addition, income information is particularly sensitive because people consider the information important. Income is a major component in defining one's social class and standing. As a result, people with low incomes may be afraid that this will reflect badly on them. People with high incomes on the other hand may be

concerned about envy or resentment, may be modest, or may be embarrassed to discuss their wealth. Also, people may be unwilling to discuss income because their true income differs from their declarations on income tax forms, loan/credit/scholarship applications, government benefit statements, and other income-based records. Efforts to convince respondents of the confidentiality of the survey reduce nonresponse to income questions (Bradburn and Sudman, 1979).

Second, the requested income information is not always known to respondents. The respondent may not be privy to the information or the exact details may not be readily accessible. Even within households income information is often, not shared by all. For example, children are probably not usually told the income of their parents. Similarly, lack of knowledge is higher among wives who are non-earners, older, and live in the South (Smith, 1985). These probably represent more traditional households in which the wage earning husband presumably is less willing to disclose his earnings to his spouse.

Even when the information is available to the respondent, it may not be readily accessible. Income questions often call for complicated and detailed information. For example, total household income on the GSS aggregates earned income from all family members, plus non-earned income from various sources (e.g. dividends, alimony, interest, government benefits, etc.) In many other cases the required information is even more demanding. Often earnings information must be reported separately for each income earner and/or for each type of income (e.g. job earnings, government benefits, business income, dividends, etc.). In some cases records are supposed to be checked to verify income amounts. It appears that the more disaggregation, precision, and documentation required, the greater the amount of missing data (Ono, 1972).Use of a show card with income categories may simplify the reporting task and help reduce non-response [Endnote 3].

The contribution of these two factors on the GSS can roughly be gauged by comparing Refusals, Don't Knows, and No Answers (usually inadvertent missing data often due to interviewer error such as a skipped question or unclear code). On the 1986-1990 surveys 49.5% of the non-response consisted of Refusals, 46.8% Don't Knows, and 3.7% No Answers.

Table 3 shows that the characteristics of those who refuse and those who don't know (DKers) are decidedly different [Endnote 4]. First, DKers were much less uncooperative with interviewers than were Refusers. Second, DKers came from households with more earners (and therefore more complex income information), but were themselves less likely to be currently employed or to have earned income during the last year. They were more likely on be dependent on other family members generating income and on those members informing them of the amounts involved [Endnote 5]. Third, DKers were less likely to be the main wage earner and in many cases were probably neither the household head nor the spouse of the head [Endnote 6]. Finally, DKers tended to be less educated and to hold less prestigious jobs when employed at all) than Refusers.

Non-response Bias Due to Missing Income Data

To judge the bias introduced by the missing data on income, we selected variables that we believed were either associated with 1) item nonresponse on income or 2) income levels. Table 4 presents the

associations with providing/not providing income data and for those providing income data their relationships to level of income. The first column reveals most variables are significantly, but weakly, related to data being missing on income. Table 5 shows the differences between the income known and missing groups and the multivariate analysis in Table 6 indicates that non-response to income is related to being uncooperative, older, not working or having an employed spouse, living with more adults, and being female.

The second column in Table 4 indicates that among those with income information the variables all have slight to substantial correlations with income level.

Comparing the two columns in Table 4 shows that non-response is consistently higher among groups with lower income (e.g. those with fewer earners, less education, lower employment levels). This makes it likely that the missing income cases actually have a lower income than the known cases and their absence creates an upward bias in reported income [Endnote 7]. The differences in Table 5 suggest however that the bias is not great since on those variables most closely related to income (e.g. number of earners, occupational prestige) the differences between those reporting income and those not reporting it are fairly small.

For those wishing to reduce the non-response bias, there are various weighting techniques, many which were designed especially for income adjustments, that can be used (David, Little Samuhel, and Triest, 1986; Kalton, Kasprzyk, and Santos, 1981; Rubln, 1983; Kalton, 1983; Ford, 1983; Turner and Lawes, 1983, Oh and Scheuren, 1980; Oh, Scheuren, and Nisselson, 1980; Huggins and Weidman, 1986a; 1986b).

Conclusion

Missing data is a far greater problem for income than for any other demographic and non-response has increased over time. However, the magnitude of the problem on the GSS is less than on many other studies. Nonresponse on income is reduced on the GSS by minimizing the complexity of the task. Techniques that might further reduce missing income data would include 1) repeating the confidentiality pledge when the question is asked 2) asking people who can't remember to either report what they do recall and/or consult records or other household members, and 3) asking for estimates from those who can't remember or aren't privy to the household income information.

Non-response bias on income does exist, but it appears to be fairly small. This in part results from the fact that bias from DKers and Refusers tend to cancel out. DKers tend to have lower incomes, while Refusers tend to have higher incomes. What bias does occur can be compensated for by appropriate weighting techniques.

Endnotes

(1) It is possible that item non-response bias might off-set an opposite bias due to sample non-coverage, survey nonresponse, or some other factor. Under this situation the bias from item non-response would create a less biased overall estimate.

(2) The GSS are full-probability, personal surveys of the adult population of the United States living in households. For complete information see Davis and Smith, 1990.

(3) Show cards presenting income categories are believed to increase response. First, they simplify the reporting process by indicating that exact reports are not needed. Second, they relieve people from having to mention dollar amounts since a letter signifying a category can be mentioned instead. Third, they increase confidentiality somewhat by avoiding respondents mentioning income levels in front of third parties (e.g. family members and visitors). However, there have apparently been no experimental studies demonstrating the utility of using show cards on income questions.

(4) Because of their small magnitude No Answers are ignored.

(5) Missing income information is more likely to come from Don't Knows when the respondent either did not earn money last year or was not employed last week and when there are more earners in a household. For example, among those working last week the % of missing cases due to Don't Knows was 28% with one wage earner, 30% with two, and 46% with three+. Among those not working last week the % Don't Know was 61% for one earner, 78% for two, and 88% for three+.

(6) We can't demonstrate this directly, but DKers tend to come from households with three or more adults and tend to be either relatively young or relatively old. We believe that many DKers may be either adult children living in the parental home or elderly parents living in their child's home.

(7) Some previous studies find that those with income missing have lower income than responders (Ono, 1972; Kalton, Kasprzyh, and Santos, 1981), while others report nonresponders having higher incomes (Oh and Scheuren, 1980; Oh, Scheuren, and Nisselson, 1981).

Table 1

Item Non-Response on Standard GSS Demographics

(1972-1990)

Variables	% Missing
Household Income	8.2
Family Relative Income when Child	1.1
Number of Earners	1.0
Number of Unrelated Household Members	0.6
Age	0.4
Party Identification	0.4
Occupation	0.4
Religion	0.3
Number of Children	0.3
Years of Schooling	0.3
Born in USA	0.3
Number of Siblings	0.2
Community Type when Child	0.2
Family Status when Child	0.1
Marital Status	0.0
Labor Force Status	0.0
Race	0.0
Sex	0.0

Table 2

Item Non-Response to Income-Related Items on Selected Surveys

A. US Surveys

Variable	Date	Survey	% Missing
Family Income	1960	Census	10.6(a)
Family Income	1970	Census	20.7
Family Income	1973	CPS	15.69(b)
Hourly Wage	1978	TSDP	10.1(c)
Quarterly Earnings	1978	ISDP	54.7
Monthly Earned Income over Four Months	1983	SIPP	15.5(d)
Family Income	1983	ANES	8.7(e)

B. Cross-national Surveys (f)

Variable	Date	Survey	% Missing
Family Income	1989	West Germany	27.6
Family Income	1989	Netherlands	26.6
Family Income	1989	Italy	11.9
Family Income	1989	Ireland	11.8
Family Income	1989	Britain	11.6
Family Income	1989	USA	8.9
Family Income	1989	Hungary	4.1

a=1960 and 1970 Census figures (Ono, 1972)
b=Current Population Survey (Oh, Scheuren, and Nisselson, 1980)
c=Hourly wages and quarterly earnings from Income Survey Development Program (Kalton, 1983)
d=Survey of Income and Program Participation (McMillen and Kasprzyk, 1985)
e=American National Election Study, calculated by author

f=From the 1989 Survey on Social Inequality of the International Social Survey Program, calculated by author

Table 3

Characteristics of Refusals and Don't Knows on Household Income

(1986-1990)

	Refusals	Don't Know
% Important/Hostile (Int. Rating)	22.6	8.0
% Most People Trustworthy	38.3	32.9
% 3+ Earners	13.0	17.5
% 3+ Adults	15.4	35.5
% Less Than 30	11.8	28.6
% Greater than 64	29.2	38.5
% Employed Money Last Year	52.9	29.4
% Earned Money Last Year	56.5	26.8
% Female	59.8	72.5
Mean Years of Schooling	12.9	10.7
Mean Occupational Prestige	42.1	34.0
Mean Age	53.3	51.5

Table 4

Associates of Missing Income Data and of Income Level

(1986-1990)

Pearson's r

Variables	Missing/Not Missing	Income Level
Number of Earners	-.036**	.487**
Employed Spouse/Respondent (Not Employed) (a)	.141**	.427**
Number of Adults	-.063**	.324**
Gender (Female)	.054**	-.148**
Age	.017**	-.150**
Race (Black)	.015	-.180**
Relative Income as Child (Above Average)	-.016**	.168**
Average Years of Schooling (b)	-.075**	.432**
Top Occupational Prestige (c)	-.039*	.422**
Int. Rating as Cooperative (Not Cooperative)	-.192**	-.102**
Trust People (No)	.029*	-.197**
Income Tax (Pay too Little)	.042**	-.161**

*=Statistically significant at .05-.01 level
**=Statistically significant at .01 level or less

a=Respondent and/or Spouse Employed vs. Neither Employed
b=Respondent's Education or Average of Respondent and Spouse
c=Occupational Prestige of Respondent or Spouse whichever Higher

Table 5

Differences in Attributes by Missing/Not Missing on Income

1986-1990 GSS

Missing	Income Reported	Income
% Black	11.8	13.6+
Mean Relative Income as Child	2.8	2.8+
Mean Top Prestige	44.2	42.2
Mean Average Years of Schooling	12.7	11.9
Number of Earners	1.5	1.3
% Employed Spouse/Resp.	75.0	53.2
Number of Adults	1.9	2.1
Mean Age	44.9	51.6
% Female	56.0	65.4
Mean Interviewer Cooperation Rating	1.2	1.6
% Trusting People	40.5	35.6
% Taxes Too High	40.2	45.8

+ = Not statistically significant at .05 level.

Table 6

Correlates of Missing/Not Missing on Income

Variables LogitRegression	OLS Regression	
	Coefficient & Prob.	Coefficient/SE
Cooperation (Hostile)	.151/.000	11.4
Number of Adults	.123/.000	7.6
Employed (Either)	.104/.000	5.6
Age	.063/.000	4.0
Gender (Female)	.029/.000	2.4
Number of Earners	-.016/.373	-0.4
Average Education	.009/.556	0.4
Occupational Prestige	-.006/.638	-0.3

r squared=0.05

n=6487

References

Bradburn, Norman M. and Sudman, Seymour, Improving Interview Method Questionnaire Design. San Francisco: Jossey-Bass, 1979.

David, Martin; Little, Roderick J.A.; Samuhel, Michael E.; and Triest, Robert K., "Alternative Methods for CPS Income Imputation," JASA, 81 (March, 1986), 29-41.

Davis, James A. and Smith, Tom W., General Social Surveys 1972-1990: Cumulative Codebook. Chicago: NORC, 1990.

Ford Barry L., "An Overview of Hot-Deck Procedures," in Incomplete Data in Sample Surveys, edited by Wilijam G. Mad Harold Nisselson, and Ingran Olkin. New York Academic Press, 1983.

Huggins, Vicki J. and Weidman, Lynn, "An Investigation of Model Based Imputation Procedures Using Data from the Income Survey Development Program," SIPP Working Paper Series No. 8603. Washington, D.C.: Bureau of the Census, 1986a.

Huggins, Vicki J. and Weidman, Lynn, "An Investigation of the Imputation of Monthly Earnings for the Survey of Income Program Participation Using Regression Models," SIPP Working Paper Series No. 8606. Washington, D.C.: Bureau of the Census, 1986b.

Kalton, Graham, Compensating for Missing Survey Data. Ann Arbor: ISR, 1983.

Kalton, Graham; Kasprzyk, Daniel; and Santos, Robert, "Issues of Nonresponse and Imputation in the Survey of Income and Program Participation," in Current Topics in Survey Sampling, edited by D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press, 1981.

Marquis, Kent H.; Marquis, Susan; and Polich, J. Michael, "Response Bias and Reliability in Sensitive Topic Surveys," JASA, 81 (June, 1986), 381-389.

McMillen, David B. and Kasprzyk, Daniel, "Item Nonresponse in the Survey of Income and Program Participation," American Statistical Association 1985 Proceedings of the Section on Survey Research Methods. Washington, D.C.: GPO, 1985.

Oh, H. Lock and Scheuren, Frederick J., "Estimating the Variance Impact of Missing CPS Income Data," American Statistical Association 1980 Proceedings of the Section on Survey Research Methods. Washington, D.C.: ASA, 1980.

Oh, H. Lock; Scheuren, Fritz; and Nisselson, Hal, "Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing CPS Income Data," American Statistical Association 1980 Proceedings of the Section on Survey Research Methods. Washington, D.C.: ASA, 1980.

Ono, Mitsuo, "Preliminary Evaluation of 1969 Money income Data Collected in the 1970 Census of Population and Housing," American Statistical Association Proceedings of the Social Statistics Section, 1972. Washington, D.C.: ASA, 1972.

Rubin, Donald B., "Imputing Income in the CPS: Comments on 'Measurement of Aggregate Labor Cost in the United States'," in The Measurement of Labor Cost, edited by J. Triplett. Chicago: University of Chicago Press 1983.

Smith, Tom W., "An Analysis of the Accuracy of Spousal Reports,"GSS Methodological Report No. 35. Chicago: NORC, 1985.

Turner, Ralph and Lawes, Murray, "Incomplete Data in the Survey of Consumer Finances," in Incomplete Data in Sample Surveys, edited by William G. Madow, Harold Nisselson, and Ingram Olkin. New York: Academic Press, 1983.