VALIDATING A SURVEY ESTIMATE - A COMPARISON OF THE GUYANA RURAL FARM HOUSEHOLD SURVEY AND INDEPENDENT RICE DATA

David J. Megill, U.S. Bureau of the Census

I. Background

While attempting to validate survey estimates, it is often difficult to obtain accurate independent estimates with which to compare the survey results. When other sources of data are found, one must carefully examine the methodology and definitions used in obtaining this information. Since both the survey and independent data have to be "validated" in this way, sometimes it is not possible to determine which data are more accurate. Even when reinterview procedures designed to provide more accurate data are used for validation, the reinterview data may still be subject to a response bias similar in direction if not magnitude to that of the original survey data. For example, a reinterview usually relies on the response to the same or similar questions using the same respondent. This paper illustrates the issues of validation and the subsequent implications for data publication, using data on rice acreage and production collected from a national sample survey in Guyana, South America.

The Guyana Rural Farm Household Survey (GRFHS) was developed jointly by the Guyana Ministries of Economic Development and Agriculture, the U.S. Agency for International Development (USAID) and the U.S. Bureau of the Census. We provided technical assistance in designing the sample and planning the survey methodology. There were two major objectives to the survey: one was to obtain socioeconomic characteristics of small and poor farm households, and another was to measure the total area and production for most crops and obtain estimates of numbers of livestock. Therefore the survey was designed to fulfill, to the extent possible, the data requirements for both objectives.

II. Survey and Sampling Methodology

The GRFHS questionnaire was designed to collect detailed information on crop acreage and production, farming technology, cost of production, and household income, with an overall reference period of one year. The sampling and analytical unit for the survey was a rural farm household which operates at least half an acre of land or owns a minimum number of animals (5 heads of cattle and/or 3 breeding sows and/or 10 sheep and goats and/or 100 heads of poultry).

The sampling frame used for the GRFHS was the Guyana National Farm Registry, which was a listing of agricultural plots by village. Since the Registry had been completed over a period of 3 years, some of the information was outdated, although there were no major changes in the agricultural sector during that time. In order to improve estimates of total crop production, the 79 farms in the Registry with 250 or more acres were identified to be included in the sample with certainty.

The villages in the Farm Registry were used as primary sampling units (PSU's) in a two- or three-stage design (except in one region where a one-stage selection was carried out using a list frame from a complete enumeration). A total of 242 PSU's were selected with probabilities proportional to estimated number of farm households. This estimated measure of size was obtained from the Registry information and adjusted using Population Census data on number of households. The larger PSU's were divided into chunks of about 75 households each, and one or two chunks were randomly selected. A listing of the households in the sample PSU's and chunks was carried out from October to December 1978. The listing sheet included screening questions to identify farm households.

At the final stage of selection, the farm households with 50 or more acres were included with certainty, and the medium size farms (2.50 to 49.99 acres) were sampled at twice the probability of the small farms (up to 2.49 acres), in order to decrease the sampling error for estimates of total acreage and production in crops. A sample of about 2800 farm households was selected from the listing, stratified by region and farm size. The survey questionnaire was administered to the sample farm households from January to March 1979.

III. Validation

A. Comparison with Independent Source

Since rice is one of the most important crops in Guyana for export revenues and local consumption, it was considered critical to validate the GRFHS estimates of total rice acreage and production before using them as a basis for policy decisions. The primary source of independent rice data was the Guyana Rice Board (GRB), which controls the marketing of rice and provides major inputs for its production. The GRB has an ongoing data collection operation for estimates of rice acreage and production by harvest season.

A preliminary comparison of the GRFHS and GRB rice estimates showed that the GRFHS estimates of total acreage and production were both more than 30% below the corresponding GRB estimates. Because of the importance of these estimates, we decided to carry out a comprehensive investigation of the sources of this discrepancy. The percentage of difference between the GRB and GRFHS estimates was similar for rice acreage and production. Since more independent data are available for area, we based our investigation on acreage of riceland.

The GRB figure for total area in riceland is 229,823 acres, while the corresponding GRFHS estimate is 148,027. However, while the GRB frame was designed to include all of the riceland in Guyana, there were certain types of agricultural operations which were excluded from the GRFHS frame, as a household survey approach was not appropriate for obtaining the corresponding data. Therefore, it is necessary to account for the riceland in these operations in the comparison between the GRB and the GRFHS estimates. The only riceland known to be excluded from the GRFHS frame was that operated by the GRB and "true" cooperatives (where the land and production are controlled jointly by the members). The GRB has 5000 acres of riceland, and only two "true" cooperatives were identified, with a total of about 82 acres of riceland, so the estimate of total area of riceland excluded from the GRFHS frame is 5082

acres. Therefore, the GRB estimate of total rice acreage corresponding to the GRFHS frame (i.e., subtracting the 5082 acres excluded from the GRFHS frame) is 224,741, and the discrepancy between the two estimates is 76,714, or 34.1% of the GRB figure. The estimated coefficient of variation for total area of riceland from the GRFHS is 5.7%, so the difference between the GRB and GRFHS estimates cannot be accounted for by sampling error alone.

B. Reconciliation of Differences

1. Evaluation of GRB Data

The first part of the reconciliation of differences involved evaluating the accuracy of the independent GRB data. Although there is no written documentation on GRB's methodology for data collection and estimation procedures, a verbal description was provided by the Director of Research and Production. Essentially, the GRB uses a "closed segment"[1] frame in which the area of riceland is identified within each village. Crop reporters in different areas are responsible for sending in updated reports on total area of riceland, area planted and area harvested. Based on this verbal description of the methodology, it appears that the GRB estimates of acreage in riceland are reasonably accurate. Their definition of riceland, "area usually used for rice" is fairly consistent with that used in the GRFHS ("acres normally used as riceland"). However, given the decline in the area of riceland which has been cultivated in the last few years, there is a possibility of a small classification inconsistency between the GRFHS and the GRB data.

The GRB estimates of total rice production are obtained from a count of the number of bags of rice harvested per acre at different sites by the GRB combines and a few private combines. The average yield per acre is then multiplied by the total acreage of rice harvested to estimate the total production. The bag count at each site should be fairly accurate, since the farmers pay for the combining per bag. However, this procedure is somewhat biased because the sites are not a probability sample. This bias is probably upward, since the combines could reach some of the better rice fields. The GRB has also made an independent estimate of rice production from the records on the amount of exports of rice produced in 1978, estimates of local consumption, amount retained for seeds, etc., which was about 5% lower than the estimate from the GRB frame. This also indicates a slight upward bias in the GRB estimate of rice production. However, since the differences between the GRB and the GRFHS estimates of total rice acreage and production are both over 30%, it is apparent that the GRFHS rice estimates do suffer from a serious downward bias.

2. Investigation of Potential Sources of Downward Bias in GRFHS Estimate

One common source of downward bias for survey estimates of totals is undercoverage. We could not identify any serious undercoverage problem associated with the implementation of the chunking (sub-segmenting of PSU's) or listing procedures. The staff of the Ministry of Economic Development involved with the GRFHS were confident that the crop reporters who carried out the listing were familiar with the boundaries of the villages and that the sample areas were canvassed fairly well. The chunking also appeared to be implemented as specified. However, an expensive study would be required to measure undercoverage objectively. There is no reliable independent estimate of the total number of farm households available with which to compare the GRFHS estimate.

One problem associated with the large farms in the GRFHS was that some of them had been divided into several parts since the Farm Registry had been completed, in which case the interviewers only obtained survey information for the part of the original land associated with the household corresponding to the original farm. The staff of MED believes that the different owners usually live in separate households and therefore have a probability of selection in the main frame. Although there is still potential bias associated with this problem, its contribution to the overall discrepancy is probably minor.

The large farm nonresponse bias may be a significant component of the discrepancy, since the nonresponse rate for the list frame of large farms (with 250 or more acres) included in the sample with certainty was 35.4%. The main reason for such a high nonresponse rate for the large farms was refusal to be interviewed, even though several attempts were made by survey personnel to collect the data. The overall nonresponse rate for the survey was only about 5.3%. This probably indicates that the large farm operators were more suspicious of government taxation and land reform policies. The estimated average area of riceland for the nonrespondent large farms is 369.9 acres (based on the Farm Registry and reinterview data), compared to an average of 203.0 acres for the respondent farms (based on the survey data). Since the average rice acreage for the nonrespondent farms is almost twice that for the respondent farms, the survey estimate of total area of riceland will suffer from a downward bias. (The weighting nonresponse adjustment factor was simply the number of valid sample units divided by the number of respondent units, by stratum). Considering the Farm Registry data for nonrespondent large farms reasonably accurate, the large farm component of the estimate of total rice acreage could be adjusted to reduce this nonresponse bias, as follows:

$$A_L = A_{LR} + A_{LN} = 10,354 + 10,356 = 20,710$$

where:

- AL → adjusted estimate of total rice acreage for large farms
- A_{LR} → total rice acreage (unweighted) for respondent large farms, from the GRFHS data
- A_{LN} → total rice acreage for nonrespondent large farms, based on Farm Registry data

The GRFHS weighted estimate of total rice acreage for large farms is 18,481.7, or 11% less than the adjusted estimate A_L, indicating the potential size of the large farm nonresponse bias. The large farm estimate A_L can be used to adjust the survey estimate of total rice acreage, as follows:

$$A_T = A_S + A_L = 129,545 + 20,710 = 150,255,$$

where:

- A_T → adjusted estimate of total rice acreage
- A_S → survey estimate of total rice acreage for farms with less than 250 acres

The difference between the GRB estimate of total area of riceland and this adjusted GRFHS

estimate is 74,486 acres, or 33.1% of the GRB figure. Therefore this adjustment for large farm nonresponse bias would only result in a slight reduction in the discrepancy.

3. Reinterview

In order to investigate whether the response error and bias for the survey estimate of total rice acreage are potentially serious, we carried out a reinterview study in two regions where rice is a predominant crop. We selected eight of the sample villages (PSU's) in West Coast Demerara and nine villages in the Mahaica area of East Coast Demerara for this study, so that each GRFHS respondent household in these villages could be reinterviewed. The set of questions from the GRFHS on the area of farmland, by tenancy and land-use (including area of riceland), was used for the reinterview so that a measure of potential bias could be obtained for farm size and rice acreage.

The training for the reinterview was carried out separately in the two regions. During the training, emphasis was given on probing and sketching the individual plots of household members in order to obtain more accurate data on area. The eight enumerators for the reinterview were crop reporters who had previously been interviewers for the GRFHS; the crop supervisors in each region helped coordinate the reinterview effort. In West Coast Demerara, 50 sample farm households were selected for the reinterview (of which six were nonresponses); in Mahaica, 52 reinterviews were assigned (of which four were nonresponses). Tables 1 through 4 summarize the results of the reinterview for total area of farmland and total area of riceland in each region, by two farm size groups (less than 50 acres and 50 or more acres, based on area in the original questionnaire). In these tables, the farm households are divided into three categories, depending on whether the area reported in the reinterview was more than, less than, or equal to that reported in the original questionnaire.

Tables 1-4 indicate that there is a potentially serious downward response bias in the GRFHS estimate of total area of riceland. Since the reinterview was fairly short and concentrated only on area, and the interviewers used sketching and probing techniques, we would expect the reinterview data to be more accurate in most cases. However, the reinterview was carried out by some of the same interviewers who worked on the original survey and is also subject to a certain amount of response error and bias, especially that from "deliberately" incorrect responses.

The tables show response error in both directions, also indicating the possible limitations of the reinterview data. Assuming that the reinterview data are more accurate than the original questionnaire data, the estimated downward "bias" for the estimates of total rice acreage is 11.1% for West Coast Demerara and 28.6% for Mahaica. It is interesting to find that in West Coast Demerara the estimated downward bias for total area of farmland (21.9%) is about twice that for riceland (11.1%); while in Mahaica the downward bias for total area is quite small (3.5%), compared to a considerably large bias for riceland (28.6%). Apparently in West Coast Demerara some respondents failed to report their total area, including riceland, in the survey, while in Mahaica the total area of farmland was reported fairly accurately, but there was a tendency for respondents to misclassify the land-use for riceland. In West Coast Demerara the bias is concentrated in the farm size group with less than 50 acres, while in Mahaica the bias is concentrated on the large farms. However, there were only two farms with 50 or more acres reinterviewed in West Coast Demerara and only six in Mahaica. It should be noted that in West Coast Demerara one farm household accounts for 45 acres of the difference in riceland, or 87.4% of the bias, and in Mahaica two large farms account for 397 acres of the difference, or 95.8% of the bias.

An attempt was also made to obtain the reinterview information for the 27 GRFHS nonrespondent large farms and 14 large farms with a considerable discrepancy between the listing and survey data on area. However, only six reinterview questionnaires were completed for the nonrespondent large farms, and seven of the large farms with an area discrepancy were reinterviewed. The latter seven reinterviews also indicated a considerable response error and downward bias for riceland on the large farms. The data are summarized in Tables 5 and 6.

For these seven large farms, the percentage of downward "bias" was very high (over 90%) for both total area and acreage of riceland, although two farms accounted for most of this discrepancy. We cannot generalize from this study of seven large farms, but it indicates the possible effect of the response bias for a few large farms on the rice estimates. For five of the nonrespondent large farms, the estimate of total area of riceland from the reinterview data was 45% lower than that from the Farm Registry data, indicating possible limitations of the reinterview data also.

4. Independent Data for Large Farms

Independent data on rice acreage were obtained for 45 large farms (33 respondents and 12 nonrespondents), from GRB records of riceland registration. The total area of riceland (for the 33 respondent farms) from the GRB records is 71.4% higher than that from the GRFHS data, also indicating a potentially serious downward response bias for the survey rice estimates. Since the farmers can only receive inputs (such as combine services and fertilizer) from GRB for the land they register, there is an incentive to report all their land. However, the GRB data pertain to farm operations, and it is possible that one operation may be divided among two or more households, although this is probably rare. Using the rice data from the GRB records for these 45 large farms and the GRFHS, reinterview or Farm Registry data for the remaining 34 large farms, the estimate of total area of riceland on large farms is 21,533. The corresponding adjusted national estimate of total area of riceland is equal to 151,079 acres. The difference between the GRB figure and this adjusted estimate is 73,662 acres, so the discrepancy is only reduced to 32.8%.

IV. Conclusions

The percentage of difference between the GRB and GRFHS estimates of total rice production is similar to that of total area of riceland, so it appears that the downward response bias is consistent for survey estimates of total rice acreage and production. (No serious undercoverage was found, although this problem is still a

possibility). The reinterviews indicated that the response bias was especially serious for large farms. One possible reason for this consistent under-reporting of area and production of rice is that some sample farm households may be responsible for more than one rice operation but failed to report the area and production for each operation. This could especially be a problem for large farms, which may divide their farm into more than one operation for tax purposes.

Given the apparently serious downward bias for the GRFHS estimates of total rice acreage and production, it was recommended that the corresponding estimates from GRB be used instead, with the slight upward bias of the GRB rice production estimates taken into account. It was not possible to adjust the survey rice estimates to account for this downward bias because it appears that the main component is an unquantified response bias (the reinterview study only indicated the potential nature and level of this bias). However, given the consistent bias for rice acreage and production, the survey ratio estimates (averages and proportions) related to rice appear quite reasonable. Several GRFHS rice estimates in the form of ratios were compared to corresponding independent estimates. The GRFHS estimate of average yield of rice per acre was exactly the same as the GRB estimate for the 1978 spring crop (17.1 bags per acre) and only 1% higher for the autumn crop (16.7 bags per acre).

Perhaps the most important conclusion to be drawn from this investigation is that it illustrates the vital need to plan nonsampling error studies into any data collection activity. This is especially important for any first-time survey effort, and is absolutely necessary in those cases where independent data on the characteristic of interest do not exist. Having such data for rice acreage and production in Guyana, we were motivated to investigate the difference between the survey and independent GRB estimates, and found a significant downward bias in the survey estimates. If we did not have such data, and lacking the nonsampling error research, the erroneous data may have been used without further verification. In many surveys time and money are invested in obtaining good estimates of sampling error. However, the Guyana study shows that this error may be minor compared to the nonsampling error and bias which are seldom investigated.

Footnote

[1] "Closed segment" methodology involves accounting for all the land within the boundaries of an area segment.

Table 1. West Coast Demerara-Difference Between Total Area of Farm Land Reported in Reinterview (R) and Original Questionnaire (Q), by Farm Size

Farm Size	Reporting Category	Number of Cases	Total Acreage of Farm Land R	Q	Difference (R-Q)	% Difference (of R)
Less than 50 acres	R>Q	18	347.07	199.14	147.93	42.4
	R<Q	21	78.23	99.65	-21.42	-27.4
	R=Q	3	14.35	14.35	-	-
	Total (net)	42	439.65	313.14	126.51	28.8
50 or more acres	R>Q	-	-	-	-	-
	R<Q	2	135.29	135.75	-0.46	-.003
	R=Q	-	-	-	-	-
	Total	2	135.39	135.75	-0.46	-.003
All Farm Households	R>Q	18	347.07	199.14	147.93	42.4
	R<Q	23	213.52	235.40	-21.88	10.2
	R=Q	3	14.35	14.35	-	-
	Total (net)	44	574.94	448.89	126.05	21.9

Table 2. West Coast Demerara-Difference Between Total Area of Riceland Reported in Reinterview (R) and Original Questionnaire (Q), by Farm Size

Farm Size	Reporting Category	Number of Cases	Total Acreage of Riceland R	Q	Difference (R-Q)	% Difference (of R)
Less than 50 acres	R>Q	12	205.08	132.07	73.01	35.6
	R<Q	10	47.50	69.00	-21.50	-45.3
	R=Q	20	76.25	76.25	-	-
	Total (net)	42	328.83	277.32	51.51	15.7
50 or more acres	R>Q	-	-	-	-	-
	R<Q	-	-	-	-	-
	R=Q	2	134.00	134.00	-	-
	Total (net)	2	134.00	134.00	-	-
All Farm Households	R>Q	12	205.08	132.07	73.01	35.6
	R<Q	10	47.50	69.00	-21.50	-45.3
	R=Q	22	210.25	210.25	-	-
	Total (net)	44	462.83	411.32	51.51	11.1

Table 3. Mahaica-Difference Between Total Area of Farm Land Reported in Reinterview (R) and Original Questionnaire (Q)

Farm Size	Reporting Category	Number of Cases	Total Acreage of Farm Land R	Q	Difference (R-Q)	% Difference (of R)
Less than 50 acres	R>Q	19	465.03	314.72	150.31	32.3
	R<Q	17	154.91	199.95	-45.04	-29.1
	R=Q	6	71.25	71.25	-	-
	Total (net)	42	691.19	585.92	105.27	15.2
50 or more acres	R>Q	3	791.00	687.50	103.50	13.1
	R<Q	3	601.75	738.00	-136.25	-22.6
	R=Q	-	-	-	-	-
	Total (net)	6	1392.75	1425.50	-32.75	-.02
All Farm Households	R>Q	22	1256.03	1002.22	253.81	20.2
	R<Q	20	756.66	937.95	-181.29	-24.0
	R=Q	6	71.25	71.25	-	-
	Total (net)	48	2083.94	2011.42	72.52	3.5

Table 4. Mahaica-Difference Between Total Area of Riceland Reported in Reinterview (R) and Original Questionnaire (Q), by Farm Size

Farm Size	Reporting Category	Number of Cases	Total Acreage of Riceland R	Q	Difference (R-Q)	% Difference (of R)
Less than 50 acres	R>Q	13	255.38	151.25	104.13	40.8
	R<Q	13	123.50	175.25	-51.75	41.9
	R=Q	16	65.50	65.50	-	-
	Total (net)	42	444.38	392.00	52.38	11.8
50 or more acres	R>Q	3	772.00	370.00	402.00	52.1
	R<Q	1	60.00	100.00	-40.00	-66.7
	R=Q	2	175.00	175.00	-	-
	Total (net)	6	1007.00	645.00	362.00	35.9
All Farm Households	R>Q	16	1027.38	521.25	506.13	49.3
	R<Q	14	183.50	275.35	-91.75	-50.0
	R=Q	18	240.50	240.50	-	-
	Total (net)	48	1451.38	1037.00	414.38	28.6

Table 5. Seven Large Farms Reinterviewed-Difference Between Total Area of Farm Land Reported in Reinterview (R) and Original Questionnaire (Q)

Reporting Category	Number of Cases	Total Acreage of Farm Land R	Q	Difference (R-Q)	% Difference (of R)
R>Q	5	1873.55	186.26	1687.29	90.1
R<Q	1	3.75	4.62	-0.87	-0.2
R=Q	1	100.00	100.00	-	-
Total (net)	7	1977.30	290.88	1686.42	85.3

Table 6. Seven Large Farms Reinterviewed-Difference Between Total Area of Riceland Reported in Reinterview (R) and Original Questionnaire (Q)

Reporting Category	Number of Cases	Total Acreage of Riceland R	Q	Difference (R-Q)	% Difference (of R)
R>Q	3	559.00	45.00	514.00	91.9
R<Q	4	3.00	107.00	-104.00	-3466.7
R=Q	-	-	-	-	-
Total (net)	7	562.00	152.00	410.00	73.0