PUERTO RICO FAMILY INCOME SURVEY

FINAL REPORT

October 30, 2000.

INTRODUCTION

Agreement Number 85-00-MOA-02 between THE DEPARTMENT OF COMMERCE, THE BUREAU OF THE CENSUS (BOC), and the PUERTO RICO DEPARTMENT OF LABOR AND HUMAN RESOURCES BUREAU OF LABOR STATISTISC (PRBLS), contains the terms and conditions for THE PUERTO RICO FAMILY INCOME SURVEY (PRFIS). The survey was conducted by ASEP, Inc. for the PRBLS during the months of February to October 2000. Appendix 1: Memorandum of Understanding.

The BOC has the requirement to provide the US Department of Education with poverty estimates for the Commonwealth of Puerto Rico. Under the stated Agreement, the necessary monetary income data to fulfill the BOC requirement were collected in the PRFIS. A survey instrument similar to that used in the US, in the Income Supplement of the March Current Population Survey (CPS), was developed in collaboration with the BOC and used to conduct the survey (Appendix 2). Based on PRFIS results, and using the poverty thresholds provided by the BOC, final estimates and standard errors were generated for the number and percent of people living in poverty in 1999, for the following groups:

- Related children age 5-17
- All people under age 18
- All people of any age

The final edited data from the PRFIS, including analytical weights, and the final sample management file that documents the total released sample of addresses and the final disposition of each sampling unit, are contained in Appendices 3, 4, 5 and 6 in Files: Houserpt, Persons, Intervie and Totreps, respectively.

This report: describes survey sample selection methodology; summarizes fieldwork, data validation and quality control processes; presents field outcome

tabulations; describes imputation and estimation procedures and reports final poverty estimates with standard errors. A more detailed description of specific aspects is included in Appendices 7 to 27. Survey materials, such as: frame construction, segmentation and sample selection protocols; examples of sample block maps; segmentation validation forms; questionnaire and control card; interviewer protocols; data entry and editing protocols and procedures; and data quality verifications, are part of the Appendices. Also included as appendix are working papers on imputation methodology and on standard error computations.

Section VII contains a summary of project successes and failures and recommendations for future Family Income Surveys.

II. SAMPLE DESIGN

The PRFIS was based on a 2-stage and, occasionally, 3-stage, stratified probability sample of households for all urban and rural Puerto Rico. The Summary Tape Files from the 1990 Census provided the block frame, since it is the most complete block-sampling frame in Puerto Rico.

STAGE 1

Construction of Sampling Frame and Selection of Blocks: with probability proportional to the number of occupied households in 1990.

Sample blocks were selected with probability proportional to the 1990 Census number of occupied houses per block in the first stage. Two strata were defined: a **regular stratum,** coded as "R", consisting of all urban and rural blocks with **5 or more occupied houses in 1990,** and a **special stratum,** coded as "N", **c**omposed of those blocks with **4 or less occupied houses,** most of which were empty in 1990. The major purpose of this method was to facilitate the field segmentation process, given that the special stratum blocks were generally large and unstructured geographical areas with very low population densities and deserved a low sampling rate and special segmentation procedures.

Regular Stratum

Measures of Size

Blocks were ordered according to five characteristics using 5 sorting keys: first key: Urban and Rural classification (urban, then rural); second key: Block Median Household Value (from lowest to highest value); third key: Census Tract (ascending); fourth key: Block Group (ascending), and fifth key: the Municipio code (ascending).

For each block, the number of 1990 occupied housing units was divided by 10 and rounded to the nearest integer. The result was assigned to each block as the measure of size, or number of assigned sampling units per block. These numbers were adjusted so

that the sum of these measures was a multiple of 150 (given a projected sample size of 1500 households from the regular stratum, to be selected in segments of an average size of 10 households).

Serial numbers were assigned to each block using the cumulative measures of size, so that each block in the regular stratum had an initial and an ending serial number. The total adjusted number of assigned units resulted in 106,500. Appendix 7: "Primera Fase: Preparación del Marco de Muestreo y Selección de la Primera Etapa de la Muestra", contains a detailed description of the primary unit selection procedures.

Fine Zones

The adjusted total number of assigned sampling units, 106,500, was divided by 150, to create 150 **fine zones of length 710**. A random number between 1 and 710 was chosen, and the length was added cumulatively until 150 units, each associated with a particular block and fine zone, were selected. The selected blocks constituted the first stage primary units from the regular stratum.

Thick Zones

The 150 fine zones defined 30 Thick Zones. The first thick zone includes fine zones 1 to 5, the second includes fine zones 6 to 10, etc. A pseudo replicate was defined selecting one fine zone from each thick zone as follows: First pseudo replicate, consisting of fine zones 1,6,11,16,21…146; Second pseudo replicate consisting of fine zones 2,7,12,17,22,27…147; etc, until 5 pseudo replicates, each consisting of 30 fine zones, were selected to compose the primary stage sample from this stratum.

Special Stratum

Measures of Size

Blocks were organized somewhat differently in the special stratum. Since many of these blocks were empty in 1990, the median household value was not used as a criterion. Four sorting keys were used: first key: Urban then Rural; second key: Census Tract (ascending); third key: Block Group (ascending); fourth key: Block Number (ascending). One sampling unit was assigned as the measure of size of each block.

The total number of blocks in the strata (6,190) was divided by 10 and rounded to the nearest integer to obtain a **length of 619.** A random number was selected between 1 and 619 and this length was added cumulatively until the block list was exhausted. The sample resulted in 10 blocks selected from the special stratum. Two blocks were assigned to each of the five pseudo replicates to complete the sample of 160 first-stage units: 150 from the Regular Stratum and 10 from the Special Stratum.

STAGE TWO both STRATA

Segmentation, Sampling Units (SU) and Selection of Segments

The first stage provided the sample of blocks, selected according to the 1990 Census number of occupied households. In the second stage, segments were created based on fieldwork identification of year 2000 occupied houses in each sample block. The segmentation process consisted of:

1. Creating as many segments, of about 10 occupied households each, as were necessary to accommodate all occupied housing units within the selected block. In population growth areas, the number of segments created was greater than the original measure of block size; and less than the original measure of size in blocks with diminishing population.

2. Listing segments within blocks, as follows: a first list numbering segments from **1 to the original number of sampling units (measure of size) previously assigned** in the first stage. If more segments were created, the additional segments were listed sequentially in a second column next to the previous list, forming a line or row; the next group of segments in a third column, etc., as needed. So **a group of one or more segments allocated to a row constituted a sampling unit**, **SU**. In this way, **the initial number of sampling units assigned to each block was preserved.** One SU, or row, was selected at random to represent the block.

3. Generating a frame for the final sampling stage and randomly selecting one segment in the selected SU (in the case of more than one segment per SU): If the number of households in a selected sampling unit containing 2 or more segments, was 13 or less, the whole sampling unit was selected with no sub sampling of segments. In other cases, due to cost restrictions, if the selected SU contained **m**, m greater than one, segments, a **third stage** was performed selecting one segment at random within the SU with probability **1/m**. A corresponding **weight of m** was assigned to the selected segment to counteract for the third stage selection. So that: if the SU contained one segment or 2 or more with less than 13 households in total, the whole SU was selected and a weight of 1 was assigned, if 2 segments, a weight of 2 was assigned, etc.

Appendix 6: FILE: Totreps, presents these weights, designated as **WGT,** for each fine zone. Tables 9 and 10 in Section IV: *Field Outcomes*, present the distribution of weights across households. Appendix 8: “Protocolo de Segmentación”, was used to train

personnel for the segmentation process. Also, trained staff from the PRBLS collaborated in this process.

For interviewing purposes the selected segment was defined in the protocols, as: all occupied households identified by the interviewer from the initial address in the segment, moving in the direction of the next segment with houses on the left side, to the last address before the first in the next segment.

III. FIELDWORK

The five major tasks related to field work were:

- Planning and development of protocols and survey instruments
- Segmentation of selected blocks and selection of sampling units
- Interviewer selection and training
- Data gathering, editing and entry, and data quality validation
- Supervision of field work related processes

Survey Protocols and Instruments

Protocols

Detailed protocols and materials were elaborated and discussed with staff to insure fieldwork consistency, reliability and effectiveness. These protocols addressed important aspects of the survey, some that have been discussed in the previous sections, such as: primary units selection; segmentation and weights; validation forms (segmentation); identification of sample households in the field; interview strategies; interviewer editing and data editing; data entry; data quality validation process; and, field work follow up.

The following Protocols and materials are related to Field Work, and are referred to in this Section:

- Appendix 2 : Cuestionario y Tarjeta Control
- Appendix 7 : Primera Fase: Preparación del Marco de Muestreo y Selección de la Primera Etapa de la Muestra
- Appendix 8 : Protocolo de Segmentación
- Appendix 9 : Formulario Apropiado
- Appendix 10 : Protocolo de la Investigación
- Appendix 11 : Observaciones al Entrevistador ;
  Protocolos para Editaje y Recibo de Cuestionarios;
  Protocolo para la Supervisión
- Appendix 12 : Digitalized Map
- Appendix 13 : Consentimiento de Participación

- Appendix 15 : Procedimiento para la Supervisión Telefónica y de Campo
- Appendix 24 : Primary Unit Sample
- Appendix 25-26 : Field Outcomes per Zone and Household
- Appendix 27 : Incentive Payments

Maps

Also, 160 digitalized block maps were prepared with the collaboration of the PRBLS and the Economic Studies Division of the PR Department of Labor and Human Resources. Programs MapInfo, PCensus and PSearch were used to generate relevant census block information and for creating the maps. Interviewers and staff were provided with these materials to facilitate the correct identification of sample blocks for segmentation, and later on, for identifying and including all valid households in the interviewing process. Appendix 12: "Digitalized Map".

Questionnaire, Control Card and Data and Interview Classification

A questionnaire similar to the one used in the US, in the March Income Supplement of the Current Population Survey (CPS), was developed in Spanish and English versions from February to April. Continuous communication with the BOC was instrumental in developing the final questionnaire adapted to the Puerto Rico context. Representatives from the PRBLS, the BOC and from ASEP met from January 31 to February 2, 2000 to discuss administrative and survey design issues. Particular attention was given in these meetings to the development of the interview questionnaire and the control card in the context of sources of income and pertinent auxiliary data relevant to Puerto Rico. Appendix 2: "Interview Questionnaire and Control Card".

The interview questionnaire was designed to collect household annual monetary income received during 1999 by categories: those defined in the US questionnaire and applicable to PR, and from additional equivalent sources in PR. The questionnaire was elaborated so that an interview was required from each household member 15 years or older. Income of a "global" nature was captured once, for example, Social Security Income received for children less than 15, was collected from the mother or in her absence, from a single father. Interest in joint accounts was collected once in the interview process. Also, if the interviewer was unable to locate a member after 3 visits, on the fourth visit a "proxy", preferably the head of household, would be interviewed. The Control Card includes a record of visits.

The control card also gathered the information on family composition necessary for defining families within households, later on used for family income determination in the estimation stage. Table 1 of the Control Card was used for this purpose. To this end, every household resident at the moment of interview was classified as "member" or not "member" of the household. One person was identified as the Head of Household (criteria for these classifications are included in the Control Card). Table 1 collects the

list of members with personal facts, as they applied at the time of the interview, on: gender, age, birth date, schooling, marital status, relationship with head of household, and "spouse", "mother", "father"- relationship with other members. (FILE: Persons-Tabla 1).

Answers to the questionnaire were distinguished as to**: "data", "missing value"**, or **"not applicable"** for data entry purposes. All applicable answers not provided by the respondent, whether the respondent did not know or didn't remember it, or refused to answer that particular question, were treated as "**missing values**". If the respondent refused to participate, the interview was classified as a **"refusal",** and not counted as an "interview". If the respondent participated in the interview and answered part of the questionnaire, and for whatever reason decided not to provide some or all of the applicable income data, the interview was counted as an **"interview"** and annual income was imputed in the estimation stage.

A household in which at least **one member participated in the interview**, was considered a **response household**. **Non response** households were classified as "**with family data**" and "**with no family data**" depending on whether Table 1 with the family composition facts was completed or not. Details of field outcomes with respect to interviews and household classification are presented in Tables 3,6,7 and 8 in Section IV: *Field Outcomes.*

Interviewer Selection and Training

Forty interviewers were initially selected. An intensive supervisor and interviewer two day training session was offered in March 30 and 31 in collaboration with the BOC. Special attention was given to the issues of:

- What is and is not monetary income received
- Who is a household member
- Identification of income sources in Puerto Rico
- Relationship to the Head of Household and among members
- Correspondence between payment amount and payment periodicity
- Special terms used in the questionnaire (ex. non incorporated and incorporated owned business)

A second training session was offered in April 7 and a Pilot Project took place during the last two weeks of April. It generated useful input that was shared with interviewers.

Thirty-five interviewers were selected for the PRFIS, out of those that participated in the training sessions and showed a higher level of competence in mock interviews. Subsequent interactive training sessions took place as questionnaires were returned and special aspects needed to be reemphasized.

Interviewer Workload

Segment assignment to interviewers was initiated on May 3 for pseudo replicate 1. An interviewer presentation letter summarizing the purpose and importance of the survey was prepared and issued. A participation consent document, "C*onsentimiento de Participación"* (Appendix 13) was also prepared to be signed by the interviewed. The document exposes the title, purpose, procedure, remuneration, duration, confidentiality and voluntary participation in the survey.

Interviewers were given 2 weeks per pseudo replicate to complete the assigned interviews for that pseudo replicate. The next replicate was distributed once they returned the previous one. Interviews were completed by the beginning of July.

TABLE 1

Workload by Interviewer by 2 Week Period

Replicate	Ave. Num. Of Interviews	Ave. Number Of Segments
1	25.40	1.5
2	23.35	1.4
3	20.95	1.2
4	27.50	1.6
5	25.90	1.6
Average	24.62	1.46

Interviewer **workload** was kept to a level we consider conducive to efficiency and data quality. As is shown in Table 1, on average an interviewer visited **1.46** segments and administered approximately **25 interviews** on a 2-week period. Table 2 shows that on the average, **13.14 households** were visited/interviewer in a two-week period.

TABLE 2 Average Workload by Interviewer/2week period

Replicate	**Number of Interviewers**	**Average Number of Houses**
1	20	13.55
2	21	12.47
3	23	10.40
4	20	14.95
5	16	14.35
AVE/Rep	20	13.14

Block Segmentation

Block segmentation was a critical component for operational zing the sample. Appendix 8: "Protocolo de Segmentación", was prepared for this phase, and reviewed in detail with supervisors. The protocol includes definitions of related concepts such as: types of blocks, block "walk", starting point, segmentation, valid household, block partition, rural area segmentation procedure, etc. It establishes the procedure for diagramming sub blocks and household structures within the selected blocks; presents rules for partitioning blocks into segments (office work) and for completing the segment form in a second field inspection, with relevant information for the interviewer. **The process of creating segments, defining the SU, selecting one SU per block, selecting a segment within the SU and assigning weights when a third stage applied, is described in detail in Section II*: Sample Design.***

Also, Appendix 9: "Formulario Apropiado", tabulates the sample blocks across pseudo replicates according to several variables, such as: measure of size, number of occupied houses; ratio of measure of size – present to 1990-, selected SU, number of segments in the selected SU, selected segment within the SU, segment weight, and others. This document was used for **validation and follow up** of the following processes: definition and selection of SU, the random selection of one segment in the SU and the assignment of weights relative to the third stage.

Data gathering, editing and entry; and data quality validation

Some of the Protocols listed at the beginning of this Section were oriented particularly to this phase of the survey, for training and quality control. For example, Appendix 10: "Protocolo para la Investigación", discusses basic research principles and issues related to conducting a survey: interviewer follow up strategies; different types of questions and approaches; interviewer editing; supervisor editing; number of visits and use of proxies. It was discussed with interviewers, staff and supervisors.

Appendix 11 contains various protocols and materials prepared and used in training for the different stages of data gathering, such as: receiving questionnaires, questionnaire editing and household classification. Appendix 15: "Procedimiento para la Supervisión Telefónica y de Campo", specifically addresses the follow up of a 20% sub sample of households to assure compliance with protocols and procedures.

The Quality Control procedures during data gathering were directed, particularly to:

- ✓ **Timely detection of possible sources of error and prevention, as to the following:**
    - Validate compliance with protocols in the classification of housing units into: valid, not valid and non response

- Validate that the interview was administered to all eligible members
- Verify the inclusion of all members in the family composition table
- Validate that sample households were not substituted or missed
- Verify that the "proxy" was selected after three visits.
- Validate the use of zeros, missing values and "skip patterns"
- Identify income terminology that may be causing confusion
- Verify compliance with the number of follow-up visits needed to complete the interviews.
- Detect potential sources of error.
- Verify incentive payment.

✓ **Monitoring of Data Quality Throughout**

The idea was to implement a quality control approach of: continuous input from the process to prompt feedback to improve the process. For example, certain aspects, like, whether Table 1 of the Control Card and the Record of Visits were completed, were examined with the interviewer as questionnaires were returned. Also, interviewers received comments based on previous replicate findings when they returned questionnaires.

A data quality systematic validation process, additional to the editing process, was implemented to monitor data quality, as soon as data entry for Replicate 1 was completed. This process generated substantial feedback for improving the interviewing and editing processes in replicates 2-5. The following aspects were examined:

- Lack of consistency among answers
- Identification of "unreasonable or outlier " amounts
- Identification and follow up of important missing values
- Correspondence between payment amount and periodicity of payment

- Identification of "double reporting".

Throughout the data entry stage, File: Intervie was examined with respect to the above points. A list of errors and "potential" errors was generated for each replicate. (Appendix 14). The identified items were followed up depending on the case, as follows:

- Verification with questionnaire to determine if a data entry or editing error had occurred
- Verification of answers, by phone with interviewed, in the case of "unreasonable or outlier" amounts
- Follow up of answers, by phone with interviewed, in the case of important missing values
- Communication with interviewer or respondent in the case of inconsistent answers.

Errors were corrected as applicable, both in the questionnaire and in the electronic files and discussed with interviewers throughout the process.

Field Supervision

Replicates were examined and edited as they were returned. As was previously mentioned, feedback was given to each interviewer prior to distributing the next set of questionnaires. It was a top priority that the supervision process would be a continuous source of information for improvement. In addition, two households from each sample segment (approximately 20% of households in 100% of segments) were followed up by phone or in the field to assure compliance with protocols. The follow up protocol, Appendix 15: "Procedimiento para la Supervision Telefónica y de Campo", was prepared and used for this purpose.

Approximately 300 households were contacted either by phone or in the field. Respondents were, in a great majority, very supportive of the survey. They corroborated a high level of compliance with protocols and procedures. Questions specified in the follow up protocol addressed aspects like: number of visits to complete the interview, family composition, reason for refusal and use of proxy, among others. Also, segments with very low initial response rate were visited and a different interviewer was assigned in the second attempt. This contributed to the low level of nonresponse and refusals, and to a low level of 6% "proxy" participation. (File: Intervie).

Appendix 24 presents the primary units selected for the sample; Appendices 25 and 26 present tables with field outcomes per fine zone and per household, respectively. Appendix 27 reports incentives per household.

Overall, our field quality control process emphasized four major elements: intensive interviewer and supervisor training in collaboration with the BOC; continuous

interaction with interviewers; continuous interactive feedback with staff during segmentation, data collection, editing and data entry stages; and, a strong data quality follow up and validation process from the beginning of the data collection stage. Moreover, as was mentioned, all segments in the sample were followed up either by phone or in the field to assure compliance with protocols and for monitoring data quality. These steps provided a systematic approach to insure high participation rates and reliable data.

IV. FIELD OUTCOMES

The selection process resulted in **1392** households of which **1298** were valid sample houses. A sample profile with respect to households, persons, interviews and families is included in this section. Appendices 3 to 6 contain the basic files related to the sample, which are:

- HOUSERPT: file containing the list of all households and addresses in the selected segments, classified according to response status and as valid or not valid.

- PERSONS: file with all persons of all ages in the sample, including household members in nonresponse households with family composition data. Table 1 with household composition data is included.

- INTERVIE: file with questionnaire data for all interviews in the sample.

- TOTREPS: includes weights generated in stage three, analytical weights, imputed annual income for interviews with missing values, family unit within the household, unique family and household number, among other variables.

A unique **CHECKCDE** number identifies each person in the household and is a primary key for relating files. The definition of the variables in FILE: Totreps is included in Appendix 16.

Household Classification

TABLE 3

Households by Type of Response

Count	VALID HOUSEHOLDS					NOT VALID	
TYPE OF HOUSEHOLD	RESPONSE			NON RESPONSE			
	All Persons Responded	Some Interviews Missing	Total Response Households	Non Response No Family Data	Non Response With family Data	Not Valid	Grand Total
REPLICA							
1	237	8	245	8	12	16	281
2	224	5	229	1	15	17	262
3	227	2	227	9	8	18	264
4	250	2	252	19	4	23	298
5	243	1	244	7	16	20	287
Total Households	**1181**	**18**	**1199**	**44**	**55**	**94**	**1392**
Percent	**84.8**	**1.3**	**86.1**	**3.2**	**4.0**	**6.8**	**100.0**
Percent of Valid	**90.1**	**1.4**	**92.4**	**3.4**	**4.2**	**Total Valid 100.0 (1298)**	

File Houserpt

Tables 3 and 4 show a **92.4%** household response rate in a sample of **1298** occupied housing units (including non response). Household non response was divided in: **3.4%** of valid households which provided no family composition data (44 households), and **4.2%** households for which these data was gathered thorough neighbors, or through household members that provided the information but would not participate in the survey (55). A total of **1199** households participated as **response households**, that is, households in which at least one person was successfully interviewed.

ClASSIFICATION OF VALID SAMPLE HOUSES n=1298

TABLE 4

Household Final Status

	HOUSEHOLDS WITH:						
	FULL INTERVIEWS			1181			
	MISSING INTERVIEWS			18	**1.4%**		
		SUB TOTAL		**1199**			
	NON RESPONSE						
		WITH TABLE	55				
		WITHOUT TABLE	44				
				99			
				7.6%			
	TOTAL VALID	TOTAL Household Sample				**1298**	**100%**
	NOT VALID				94	**6.8%**	
		TOTAL HOUSEHOLDS					**1392 100%**

TABLE 5

Households by Strata and by House Number within Segment

House Number in Segment	*STRATA* ***N***	***R***	***Total HOUSES***
01	7	147	154
02	4	147	151
03	3	146	149
04	3	145	148
05	2	142	144
06	1	142	143
07	1	138	139
08	1	130	131
09	1	108	109
10		83	83
11		27	27
12		12	12
13		2	2
Total Houses	**23**	**1369**	**1392**

File Houserpt N= Special Stratum R=Regular Stratum

BY PERSONS

TABLE 6 ***ALL PERSONS IN HIS/FIS SAMPLE (including non response with family data) and INCLUDING < 15 YR OLD – Type of Interview***

REPLICATE	All Income Reported*	Partial Income Missing*	All Income Missing*	Non Interview (Refusal)	Persons	Non Resp With Fam. Data>=15	Total Persons
1	631	18	7	15	671	23	694
2	636	30	5	7	678	26	704
3	630	7	2	2	641	14	655
4	674	5	4	2	685	7	692
5	675	8		2	685	37	722
Grand Total	3246	68	18	28	3360	107	3467
Persons < 15	759**	0	0	0	759	0	759
Persons >= 15	2487	68	18	28	**2601**	107	**2708**

File Persons
**applicable income*
***all children <15 included in this category for presentation*

Interviewers classified a household after complying with the number of required visits and selecting a proxy if needed. Tables 6 and 7 classify all persons according to type of interview. All children less than 15 years old appear in Table 6 classified as "all income reported" solely for presentation purposes. A total of **2,708** persons were eligible **members 15 years or older**, out of 3,467 household members of all ages, including nonresponse households with family data (of all, 925 were children 17 or younger, 759 were less than 15). Of the eligible members, **2,601** were members in the **1,199** response households that participated in the survey.

TABLE 7 *ELIGIBLE HOUSEHOLD MEMBERS >=15*

	SAMPLE INTERVIEWS					**NON RESPONSE**	
ALL REPL	All Income Reportd *	Partial Income Missing *	All Income Missing*	Non Intervie Refusal	Total	Non Resp WithFam Data	Grand Total
Persons> =15	2487	68	18	28	**2601**	107	**2708**
Percent	91.8	2.5	0.66	1.03	**96.0**	**4.0**	**100.0**
Percent	95.6	2.6	0.69	**1.08**	**100.0**		

File Persons

Table 7 shows that the refusal rate within response households was **1.08%** (Non Interview). The number of eligible persons identified in nonresponse households with family data constituted **4.0%** of the eligible sample, 107 persons. Table 4 shows that 55 households belong to this group, resulting in a household average of 1.95 eligible members for non response households with family composition data, and an average of 2.29 persons per household if we include children less than 15 years old in those households (19 children). Overall, the average number of household members was **2.77** persons per household. This suggests the possibility of an **undercount** of people, particularly at the lower age groups, both in non response houses for which family data was collected mostly from neighbors, possibly explaining the undercount, and overall.

Table 8 presents the distribution of sample persons >=15 by pseudo replicate.

TABLE 8 *PERSONS >=15*

REPLICA	TYPE OF INTERVIEW					
	All Income Reported*	Partial Income*	All Income Missing*	Non Intervie Refusal	Non Resp With Fam Data	**TOTAL**
1	496	18	7	15	23	**559**
2	468	30	5	7	26	**536**
3	489	7	2	2	14	**514**
4	528	5	4	2	7	**546**
5	506	8	0	2	37	**553**
TOTAL	2,487	68	18	28	107	**2,708**
Percent	**91.8**	**2.5**	**.66**	**1.03**	**4.0**	**100.0**

File Persons

After applying the third stage weights, the equivalent sample size for all ages and the distribution by pseudo replicate, including non response households with family composition data, is as shown in Table 9:

TABLE 9 *PERSONS ALL AGES: EQUIVALENT WEIGHTED SAMPLE*
Including Non response Households with Family Data

REPLICATE	WGT EQUIVALENT SAMPLE (weighted by third stage sampling weight)			
	1	2	3	**TOTAL**
1	543	302		**845**
2	407	424	255	**1086**
3	467	376		**843**
4	527	330		**857**
5	547	350		**897**
	2491	**1782**	**255**	**4528**

FILE: Persons
**applicable income*

The weighted household sample is:

TABLE 10

WEIGHTED SAMPLE OF HOUSEHOLDS
WEIGHTS

HOUSEHOLDS	1	2	2	3	3	3	TOTAL
Frequency With non response with family data	**896**	**326**	**326**	**34**	**34**	**34**	**1650**
Frequency With total non response	**922**	**344**	**344**	**34**	**34**	**34**	**1712**

So that the weighted sample size, adjusting for stage 3 selection of segments, resulted in the equivalent of 1712 **households**. Given a projected sample size of 1600 households according to 1990 Census data and an increase in the number of households in the last 10 years, the weighted sample size suggests the possibility of a **household undercount** of about 5% to 8%.

TABLE 11

	PERSONS >= 15 BY FAMILY UNIT IN HOUSEHOLD (With non response with family data)					
	Count of FAMNUM	FAMNUM				
	REPL	PRIMARY FAM.	SECONDARY FAM.	Grand Total		
	1	554	5	559		
	2	534	2	536		
	3	511	4	515		
	4	540	6	546		
	5	545	8	553		
	Grand Total	2684	25	2709		

File Totreps

Table 11 shows that a great majority of households in the sample are one family household. Data from FILE: Persons shows that only two minors, <= 17 years old, were members of a secondary family in the household.

BY FAMILIES

Descriptive Statistics: Count
Families, including nonresponse households with family data

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Count	1279	2.7107	2.0000	2.6247	1.4527	0.0406

Variable	Minimum	Maximum	Q1	Q3
Count	1.0000	11.0000	2.0000	4.0000

Applying the BOC family definition resulted in **1279 families** in 1254 households (Table 3) with an average of 2.71 members per family, including nonresponse households with family data. Again this statistics suggest a member **undercount** since 2.71 seems somewhat low for the average number of members per family compared to projections from the 1990 Census Data, which to our estimation should be closer to 3 to 3.15 members per household.

V. IMPUTATIONS METHODOLOGY

The analysis **classified interviews in five TYPES or "salidas"** for the purpose of income imputation, as follows:

INTERVIEW TYPE ("salida")

- 0 = all applicable income was reported by respondent, **no imputation was necessary**
- 1 = **some, but not all**, applicable income values were missing, **imputation required**
- 2 = **all** applicable income values missing, **imputation required**
- 3 = household member refusal to participate**, imputation required**
- 4 = non response household with known family composition data, **imputation required**.
- 5 = non response household with no family composition data, **imputation required.**

A **regression model approach** was used for imputing annual income for TYPES 1 to 4. Interviews (>= 15 yrs old) classified as **Type 0**, **Regular Stratum**, were used for fitting a regression model in the following manner:

Interviews were classified initially in homogeneous groups according to 6 coded variables: 5 categories of age, 2 of gender, 5 of school level, 3 of marital status, 3 of fine zone and 4 of household value. Various model were considered and tested. Groups with 0 average annual income were identified as outliers and eliminated from the model estimation process.

Table 9 defines some of the coded variables that were tested in the process.

edad	cedad	sexo	csexo	escol	cescola	estadocv	cestacv
<= 18	1	F	1	00-06	1	casado	1
19-25	2	M	2	07-20	2	nunca cas.	2
26-35	3			21-23	3	otros	3
36-55	4			24-más	4		
56more	5						

valor vv(v1170)	cvalorvv
01-02	1
03-04	2
05-06	3
07-08	4
09-más	5

The following output relates to the model with independent variables fine zone, age, gender, and marital status.

Multiple Regression

Dep. var.: LOG(PROM) SELECT PROM > 300

Ind. vars.: IND CZON
IND CEDAD
IND CESCO
IND CSEXO
IND CECIVIL

(Model)

Weights: CASOS

Constant: Yes Vertical bars: No Conf. level: 95

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	FRatio	Pvalue
Model	278.482	10	27.8482	16.7429	.0000
Error	136.389	82	1.66328		
Total (Corr.)	414.871	92			

Rsquared = 0.671249 Stnd. error of est. = 1.28968
Rsquared (Adj. for d.f.) = 0.631158 DurbinWatson statistic = 1.68858

Further ANOVA for Variables in the Order Fitted

Source	Sum of Squares	DF	Mean Sq.	FRatio	Pvalue
IND CZON	10.164803	2	5.082401	3.06	.0525
IND CEDAD	117.296424	2	58.648212	35.26	.0000
IND CESCO	90.543183	3	30.181061	18.15	.0000
IND CSEXO	45.375578	1	45.375578	27.28	.0000
IND CECIVIL	15.101919	2	7.550959	4.54	.0135
Model	278.481906	10			

This analysis showed that variable IND CZON was not strictly significant at the .05 level (the only variable that was not) and, hence, that the model could be improved.

The best fit resulted in a weighted regression, using as weights the number of observations in the group. The **model regresses the natural log of the average annual income of each homogeneous group on indicator independent variables, age, gender and schooling.** Examination of the model shows that close to 2/3 of the total variation is explained. Residual analysis and testing for assumptions showed no significant departures from major assumptions, such as: no-multicollinearity, equal variances and normality.

The regression analysis showed that the three indicator variables define groups with less variation among individuals within the groups as compared to the overall sample. As a result, a regression model was used for imputing missing income values of TYPES 1-3 and for non response households that provided family composition data in the interview process, TYPE 4, as a means for reducing the potential non response bias. In general the procedure was as follows:

- TYPE 1 (**some** applicable income missing): the regression model was applied and the annual income estimate was compared to the sum of income amounts reported during the interview. If the model estimate exceeded the sum of reported income, the model estimate was assigned as annual income for that member.
- TYPE 2, 3 and 4: (**all** applicable income amounts missing): the model estimate was used as annual income for that member.

The following output presents the fitted model that was used.

1-Model fitting results for: LOG(INGPROM) SELECT INGPROM >0

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	4.781764	0.32809	14.5745	0.0000
IND CEDAD	1.734832	0.332431	5.2186	0.0000

Independent variable	coefficient	std. error	t-value	sig.level
IND CEDAD	3.062294	0.310666	9.8572	0.0000
IND CEDAD	3.005175	0.289475	10.3815	0.0000
IND CEDAD	3.40609	0.295877	11.5118	0.0000
IND CSEXO	0.727878	0.130645	5.5714	0.0000
IND CESCOLA	0.429262	0.201286	2.1326	0.0346
IND CESCOLA	0.93137	0.249181	3.7377	0.0003
IND CESCOLA	1.560002	0.252775	6.1715	0.0000

R-SQ. (ADJ.) = 0.6265 SE= 1.333657 MAE= 0.992116 DurbWat= 2.116
155 observations fitted, forecast(s) computed for 15 missing val. of dep. var.

2-Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	473.659	8	59.2073	33.2880	.0000
Error	259.681	146	1.77864		
Total (Corr.)	733.340	154			

R-squared = 0.645892 Stnd. error of est. = 1.33366
R-squared (Adj. for d.f.) = 0.626489 Durbin-Watson statistic = 2.11567

3-Further ANOVA for Variables in the Order Fitted

Source	Sum of Squares	DF	Mean Sq.	F-Ratio	P-value
IND CEDAD	333.086132	4	83.271533	46.82	.0000
IND CSEXO	58.176632	1	58.176632	32.71	.0000
IND CESCOLA	82.395752	3	27.465251	15.44	.0000
Model	473.658515	8			

4-Residual Summary

Number of observations = 155 (15 missing values excluded)
Residual average = -0.0111346
Residual variance = 1.77864
Residual standard error = 1.33366

Coeff. of skewness = -0.345311 standardized value = -1.7551
Coeff. of kurtosis = 0.345538 standardized value = 0.878124

Durbin-Watson statistic = 2.11567

Appendices 17 and 18 contain additional details on the regression fit. Table 6 and 7 in Section IV: *Field Outcomes*, classify households according to the type of interview and, hence, to the type of imputation.

Imputation of nonresponse households – persons and income

A TYPE 5 nonresponse household was imputed from a TYPE **0** household randomly selected from households with similar socioeconomic characteristics, particularly house value, in the same segment. The family composition and the annual income per member of TYPE 0 household substituted the missing non response household data. Table 4 the in Section IV: *Field Outcomes* summarizes non response.

Imputation of Indicator Variables used as predictors

Indicator variables needed as predictors in the regression model had to be imputed for a small number of cases, due to missing information about the member in the family composition data. To this end, socio economic data collected in the questionnaire, for the household and for other household members, was analyzed to obtain "common sense" reasonable estimates for the missing indicator variables.

VI. ESTIMATION, WEIGHTS AND STANDARD ERRORS

Household members were classified by families according to the BOC Family definition, based on the information provided in the family composition data. The documents: Appendix 19: "Definición de Familia y de Niños Emparentados" and "Proceso de Creación de Familias", were prepared for operationalizing the definitions of "family" and "child", and for identifying families within households. Using poverty thresholds provided by the BOC (Appendix 20), families were classified as "poor" or "not poor". Individuals within families were classified with respect to poverty according to the family classification.

To compensate for sampling frame imperfections and possible household and "member" undercount, the choice of weights was crucial.

Weights

Third Stage Weights: WGT

Population growth areas generated more segments than the original measure of size assigned to the block. This meant that in those blocks the sampling unit, after organizing the segments so that the original number of assigned sampling units in the block was maintained, consisted of **more than one segment** of approximately 10 households each. In some of these cases, as was explained in detail in **Section II: *Sample Design;*** one segment was selected at random from within the selected sampling unit. To counteract for this third stage, weights designated **WGT** were applied to the data at the segment level. Appendix 6: File Totreps, presents these weights in column WGT. Tables 12 and 13 summarize the distribution of these weights.

TABLE 12

	WGT: All Persons including Non response Households with family data			
REPL	1	2	3	Grand Total
1	543	151		694
2	407	212	85	704
3	467	188		655
4	527	165		692
5	547	175		722
Total	2491	891	85	3467

TABLE 13

HOUSEHOLDS BY WEIGHT (stage 3 sub sampling)
Including Total Non Response

	WEIGHTS			
	1	2	3	TOTAL
HOUSEHOLDS With Non Response family data	896	326	34	1256
PERCENT	71.3	26.0	2.7	100.0
With Total Non Response	921	344	34	1299
Total	70.9	26.5	2.6	100.0

File Totreps

Table 13 shows that about 30% of households resulted from a subselection process of segments within sampling units.

Sampling Plan Projection Weights

As was explained in Section II: *Sample Design,* one out of every **710** sampling units was selected from each fine zone in the primary sampling stage in the **Regular Stratum, for a projection weight of 710**; and, one out of every **619 whole blocks** in the **Special Stratum** (see Section II on *Sample Design*) **for a projection weight** in this case **of 619**. We consolidated both weights, third stage weight, **WGT, and the projection weights in a second column of weights designated WGT2 in File Totreps.**

The projection of the sample according to its natural plan coincided with the analysis of field outcomes, which suggests a possible undercount both of households of about 5% - 8%, and of household members (a sample average of 2.77 members per household compared to 3 – 3.15, a more realistic figure from different sources). The combined effect of both types of undercount resulted in about a 14% underestimation of the population. **Section IV** presents findings with respect to the **"undercounts**".

Analytical Weights WGT3

Clearly an analytical system of weighting was needed. For this purpose we first considered the PRBLS data (from the labor force survey) which combines results from 12 of its most recent samples. This data is reliable enough to obtain reasonable estimates of the percentage of population in age groups and by region (12 regions). We then used the BOC projections of PR population totals for year 2,000 by age groups. **Combining both sets of data we estimated the 2,000 population totals by age group and by each of the 12 administrative regions of the PRBLS.** This is a 4 by 12 matrix with 48 cells.

In each of these 48 cells the weighted (by **WGT,** the third stage weights) number of persons in the sample was determined and compared to the estimated population totals previously determined. In this way an analytical weight was assigned to every person in the cell so that the projected totals would coincide with the estimated totals. The **combined effect** of the third stage weights and the analytical weights was calculated and inserted in a column as **WGT3**. The entry in any cross-classification of the data is then obtained by summing **WGT3**.

Estimates and* Standard *Errors:

Estimates and standard errors were calculated by using a ratio model. Let $\mathbf{n}_j$ be the number of persons within an age group $\mathbf{j}$ in one of any of the 12 PRBLS administrative **regions,** classified in a state of poverty, and let $\mathbf{d}_j$ be the population total within that age group as estimated by external sources and already described. Then:

$$R_j = \frac{\sum n_j}{\sum d_j}$$

is the ratio estimate of the proportion of person within that age group $\mathbf{j}$ in state of poverty in PR. The standard errors were then calculated using the well known formulas for this type of estimate:

$$s_{Rj} = \left(\frac{1}{\bar{d}} \right) \sqrt{\frac{\sum (nj - Rjdj)^2}{n(n-1)}}$$

The model assumes that $\eta = R\delta + \sqrt{\delta}\varepsilon$ from which the least square estimator of $\mathbf{R}_j$ is the one given above, a well known result. Appendix 23: "Estimación y Cómputos de Error Estándar", illustrates both, computations for the ratio estimate and for the standard error, as well as graphs showing the plausibility of this model.

Final estimates with standard errors are presented in Table 14.

TABLE 14

1999 DATA

Number of Persons in Poverty Families
By Age Group

		POVERFAM			
AGEGRP	**NO**	**YES***	**TOTAL**	**%YES***	**ST. ERR.**
00-04	**132,576**	**174,571**	**307,147**	**56.84**	**0.0251**
05-17	**369,981**	**447,687**	**817,669**	**54.75**	**0.0308**
00-17	**502,557**	**622,258**	**1,124,816**	**55.32**	**0.0258**
18-64	**1,332,599**	**1,051,280**	**2,383,879**	**44.10**	**0.0276**
65-99	**179,182**	**227,920**	**407,102**	**55.99**	**0.0484**
All Ages	**2,014,338**	**1,901,458**	**3,915,797**	**48.56**	**0.0256**

* the YES column means families in poverty

Also, Appendix 22: "Distribución del Ingreso Anual de la Familia", presents the family annual monetary income distribution for 1999, based on the PRFIS results.

VII. RECOMMENDATIONS

The PRFIS has many strengths and some weaknesses. As in any process that will be repeated in the future, a priority is to identify both, so that future performances benefit from the experience. To such end, we highlight the most important aspects in terms of strengths and weaknesses and, whenever possible offer concrete recommendations.

Questionnaire

In general the **questionnaire was an effective instrument** both from the perspective of the interviewer and the interviewed. However, improvements can be made with respect to various items, such as:

1. Question 1300 on Interest Income and question 950 on Dividend Income, had a different format for gathering the information compared to other

questions in the questionnaire. They were also associated with a higher incidence of data error. The format should be revised.

2. The Record of Visits in the Control Card proved not as effective as we had anticipated in gathering useful information about the process. In general, the format should be revised.

3. Some auxiliary questions, like for example, the type of Industry in which a person worked in 1999 and household value, proved very useful in the imputation stage. More emphasis should be placed in obtaining from the respondent this type of information. The whole aspect of auxiliary questions should be evaluated to determine which are the essential questions and the way to present them, so as to facilitate data gathering and entry, but, at the same time, controlling the questionnaire length.

Interviewers and Training

It is our appreciation that **the group of interviewers was of superior qualifications and one of the survey strengths**. Interviewers were supportive, concerned and showed exceptional commitment to the survey. Their workload was kept at very reasonable levels, which, in turn, was conducive to an effective and reliable data-gathering phase. Also, field supervision corroborated their very adequate compliance with protocols and procedures.

Although intensive trainings were offered, experience showed that some concepts were more difficult to grasp and implement in the field, and **require "reminder sessions**" after the initial training. This seems necessary in order to avoid that interviewers, during the first set of interviews, modify the way to report answers: from the way they were advised during training to one influenced by the different formats in which the interviewed reports the data. Also, mock interviews were very useful and a longer session should be programmed during training

The use of pseudo replicates

The use of pseudo replicates, as part of the sample design was **a definite plus** from the perspective of monitoring data quality. As soon as Replicate 1 was received preliminary estimates were generated and File: Intervie was examined in detail, as is explained in Section III: *FieldWork.* Various important errors in interpretation and in the administration of the questionnaire were detected then, as well as their impact on poverty estimates. Corrective action was taken and feedback was provided for the next replicates. This process was continued as other replicates were returned. This also provided the opportunity to validate the information generated from the sample as compared to similar information about PR from external sources.

Field Supervision

Phone supervision was **very effective** in corroborating certain aspects, such as: corroborating that the interview took place, following up missing values, corroborating "outlier" amounts and inconsistencies in answers, among others. However, supervision in the field **could be strengthened** to validate other important aspects, such as: confirming that all households in the segment were considered; validating that all members were included; corroborating classification of households in valid-non valid. Sample data for the PRFIS suggests a possible household and member **undercount.** This aspect needs to be analyzed more thoroughly, if so, to determine possible causes and ways to prevent it in future surveys.

Data Entry and File Management

The data entry and file management system used is one aspect that would **need special attention**. A higher priority must be given to implementing a system of electronic files and data entry, that:

- Generates periodic reports or "tallies" from the electronic files to be compared to the "paper" reports with respect to: number of interviews entered, number of persons, minors, household classification, interview classification, among other variables.

Imputations

Even though non response was small, about 8%, the regression model showed that the potential bias due to non response could be reduced by using the model to impute annual income. We believe that the imputation phase **proved successful** and should be part of the data processing for the PRFIS.

Estimation and Analytical Weights

The use of analytical weights **was crucial and effective** due to the possibility of undercount from two sources. Also, because the poverty proportions vary, by what we consider a non trivial amount, for some age groups, estimating poverty levels for all ages would suffer if age group representation in the sample does not conform to population totals.