Russia Longitudinal Monitoring Survey (RLMS)
Sample Attrition, Replenishment, and Weighting in Rounds V-VII

Steven G. Heeringa, Director
Survey Design and Analysis Unit
Institute for Social Research, University of Michigan
Ann Arbor, MI 48106-1248

March 20, 1997

1. CROSS-SECTIONAL AND LONGITUDINAL DESIGN AND ANALYSIS

Data from the RLMS may be used in two types of analyses.

A. Repeated Cross-Section Analysis. As its name implies, the RLMS is a longitudinal study of populations of dwelling units. Rounds V-VII are designed to provide a repeated cross-section sampling. Barring the construction of major new housing structures, renewed contact with a fixed national probability sample of dwelling units provides high coverage cross-sectional representation. The repeat visit at each round to a static sample of dwelling units also introduces a correlation between successive samples that leads to improved efficiency in longitudinal analyses comparing aggregate statistics.

The repeated cross-section design is far and away the simplest alternative for the RLMS. The sampling is cost efficient, easy to maintain, and easy to update when needed. The design supports both efficient cross-sectional and aggregate longitudinal analyses of change in the Russian household population. Updates to the sample, including a full replenishment of the probability sample of dwelling units, will not seriously disrupt the longitudinal data series.

B. Longitudinal or "Panel" Analysis. The primary disadvantage of a repeated cross-section design is that it does not enable micro-level analysis of longitudinal change at the household or individual level. The exception is the potential to link households and individuals who remain in the original dwelling unit over time, but such a "panel" may be vulnerable to selection bias when reasons for moving are correlated with the dependent variable of interest (see 2.B. below).

A true panel design in which sample households and individuals are followed and interviewed at each wave would be preferred if the sole purpose of the RLMS were to study micro-level change. The original sampling plan for Rounds V-VII did not call for households to be followed if they moved from the Round V sample dwelling unit. Likewise, individual household members who moved away were not to be followed. At each round, the RLMS interview was completed with the household and its members in the original sample dwelling unit. Consequently, the RLMS is not a true panel design, although Round VII departed from the original protocol and followed some households and individuals who moved.[1]

[1] True panels are costly to maintain. It takes considerable effort to trace and interview movers in order to avoid attrition bias. Experience at ISR with the Panel Study of Income Dynamics (PSID) and the Health and Retirement Survey (HRS) shows that maintaining true panels of households and affiliated individuals involves a tremendous burden of data management and cleaning. This is particularly true as the number of rounds of data collection grows. Two hybrid designs that combine elements of the repeated cross-section and pure panel designs are the split panel and the rotating panel. See Leslie Kish, *Statistical Designs for Research*, New York: John Wiley & Sons, 1987.

2. SAMPLE ATTRITION

The first question is the nature of attrition in the RLMS samples and its impact on cross-sectional and longitudinal analysis of the data.

A. Attrition Effects on the Analysis of the Repeated Cross-Section Data. Sample attrition due to nonresponse cannot be avoided. Table 1 summarizes RLMS Round V interview completion rates for the original sample of dwelling units in the eight regions that comprise the survey population. These are not response rates; each denominator includes dwelling units that were vacant or uninhabitable at the time of the Round V interviews. Overall, interviews were completed in 84.3% of the original national probability sample of n=4718 dwelling units.

Table 1: RLMS Round V Interview Completion Rates*

Region	n	Dwelling Interview (%)
Moscow/St. Petersburg	686	60.2
North/Northwestern	319	88.7
Central/Central Black Earth	923	84.8
Volga/Viask/Volga Basin	770	89.4
North Caucuses	538	87.6
Urals	619	91.0
Western Siberia	416	92.6
Eastern Siberia/Far East	447	87.0
Total	4718	84.3
St. Petersburg	*222*	*67.1*
Moscow	*464*	*56.9*

* Including vacancy, no contact, refusal.

Interview completion rates outside St. Petersburg, Moscow City, and Moscow Oblast range from 84.8% in the combined Central/Central Black Earth region to 92.6% in Western Siberia. Rates in the highly urban Moscow/St. Petersburg region are much lower. In part, these rates may reflect higher vacancy rates in metropolitan areas, but clearly lower household contact and response rates also come into play. Lower rates in Moscow and St. Petersburg were anticipated at the design stage, and initial allocations to these strata were increased to offset expected losses from refusal and noncontact. This is one form of what we might call "designing for nonresponse." The over-sampling strategy is beneficial in that it means reduced variability in the final analysis weights (due to the offset in the product of higher sample selection probability and lower response propensity); however, over-sampling eliminates the potential for bias only if attrition is occurring at random within the final weighting adjustment cells.

If independent samples were developed for each round of the repeated cross-section design, attrition in one round would be independent of (although possibly similar in nature to) that in other rounds. However, since the RLMS uses a static sample of dwellings across multiple rounds, the impact of nonresponse and attrition is the net effect of several factors. Round V attrition bias can arise only from differential nonresponse and noncontact for subclasses of households that occupy the original sample of dwelling units. The potential for nonresponse bias in cross-sectional analysis or contrasts involving the Rounds VI and VII data is a complex function of: (1) initial nonresponse in Round V; (2) net difference in characteristics of households and individuals who move out of or into sample dwellings; (3) nonresponse on the part of old

households continuing to reside in sample dwelling units; and (4) nonresponse on the part of new households currently living in sample dwelling units.

Time did not permit analysis of each of these factors. Instead, I performed several simple analyses of the net effect of household turnover and nonresponse on the marginal sample distributions (unweighted) of population characteristics that should not change significantly over time.

Table 2 compares the unweighted distribution of the Round V-VII interview households by region, settlement type, characteristics of household head, and household size. The general observation is that the combined influence of nonresponse attrition and household turnover does not seriously distort the geographic distribution of the sample or its size or household-head characteristics. The distributions for the geographic variables indicate that, between Round V and Round VII, there is a decline in the nominal representation of households in the Moscow/St. Petersburg region, reflected in a decline in the proportion of sample households from the urban domain. Households with a male head aged 18-59 may be subject to slightly higher than average attrition/net loss in replacement. If we focus only on these characteristics, the problem is not serious.

Table 2: Net Attrition/Recruitment Effect on Cross-Sectional Composition of Household Sample

Subpopulation	Percent by Category Round V	Round VI	Round VII*
REGION			
Moscow/St. Petersburg	10.4	9.2	8.5
North/Northwestern	7.1	7.2	7.3
Central/Central Black Earth	19.7	19.4	20.1
Volga/Viask/Volga Basin	17.3	17.6	17.9
North Caucuses	11.8	12.0	12.2
Urals	14.2	14.8	14.7
Western Siberia	9.7	9.8	9.4
Eastern Siberia/Far East	9.8	10.2	10.0
SETTLEMENT TYPE			
Urban	70.2	69.3	68.4
PTG	5.4	5.6	5.8
Rural	24.4	25.1	25.8
HOUSEHOLD HEAD			
Older child (7-18)	0.1	<0.1	<0.1
Male (18-59)	64.8	63.6	63.2
Female (18-54)	10.8	11.2	11.7
Male (60+)	11.6	11.8	11.9
Female (55+)	12.7	13.4	13.2
HOUSEHOLD SIZE			
1	17.6	18.7	19.0
2	26.9	26.1	26.6
3	23.8	23.7	24.0
4	21.0	20.0	19.7
5	7.0	7.6	6.6
6+	3.8	3.9	4.1

* Including households followed to new residences.

Table 3 gives a similar comparison of the unweighted marginal frequencies for individual sample members interviewed in Rounds V-VII. Again, the combined effects of attrition and change in dwelling unit occupants result in a net decline across rounds in the proportion of sample individuals from the Moscow/St.Petersburg region and an associated decline between Rounds V and VII in the percent of sample individuals from urban areas. We also find a modest decline in the proportion of males aged 0-19 between Rounds V and VII.

In summary, the net effect of nonresponse attrition and change in dwelling unit occupants across rounds on the marginal characteristics of the observed cross-sectional samples is modest. Loss in nominal "sample share" between Rounds V and VII is greatest for residents of Moscow/St. Petersburg--a loss in representation that is readily corrected with the combined sample selection/nonresponse adjustment factors that have been computed for each round. It is important to note that the simple analysis described here cannot demonstrate that no uncorrected attrition bias remains. The potential for uncorrected nonresponse bias can be specific to the dependent variable under study. Nevertheless, it appears that, with the nonresponse and post-stratification adjustments developed by Michael Swafford, the potential for serious attrition bias in repeated cross-section analysis is small.

Table 3: Net Attrition/Recruitment Effect on Cross-Sectional Composition of Individual Sample

	Percent by Category					
Subpopulation	Round V		Round VI		Round VII*	
REGION						
Moscow/St. Petersburg	10.5		9.0		8.0	
North/Northwestern	7.2		7.2		7.0	
Central/Central Black Earth	18.1		17.8		18.6	
Volga/Viask/Volga Basin	17.0		17.3		17.6	
North Caucuses	13.4		13.9		14.1	
Urals	14.4		14.9		14.7	
Western Siberia	9.9		9.8		9.7	
Eastern Siberia/Far East	9.6		10.1		10.2	
SETTLEMENT TYPE						
Urban	69.3		68.2		66.8	
PTG	5.5		5.7		6.2	
Rural	25.2		26.0		27.0	
AGE GROUP/SEX	M	F	M	F	M	F
0-19	14.5	14.0	14.3	14.0	14.0	14.0
20-39	13.9	15.6	13.6	15.3	13.6	15.3
40-59	11.1	13.6	11.4	13.6	11.3	13.7
60-79	5.5	9.5	5.5	9.8	5.5	10.2
80+	0.4	1.8	0.4	1.9	0.5	1.9

* Including individuals followed to new residences.

B. Attrition Effects on Simulated "Pure Panel" Analysis. The intent behind the RLMS design is that data be analyzed as repeated cross-sections of the Russian population. An interesting question is, "How misleading would it be to conduct pure panel analysis of households and individuals observed in Rounds V and VI or in Rounds V-VII?" The obvious problem is that by definition analysis can include only households and individuals who continue to reside in the original sample dwelling units and who participate in two or three consecutive rounds of the study.

Table 4: Attrition Effects for Round V Household Panel, Round V Characteristics for Retained Sample

Subpopulation	Percent by Category Round V Panel	Round VI Panel	Round VII Panel*
REGION			
Moscow/St. Petersburg	10.4	8.4	7.5
North/Northwestern	7.1	7.4	7.3
Central/Central Black Earth	19.7	20.1	20.6
Volga/Viask/Volga Basin	17.3	18.3	18.8
North Caucuses	11.8	11.8	12.2
Urals	14.2	14.8	15.0
Western Siberia	9.7	9.6	9.6
Eastern Siberia/Far East	9.8	9.6	9.0
SETTLEMENT TYPE			
Urban	70.2	67.2	65.7
PTG	5.4	5.6	5.6
Rural	24.4	27.2	28.8
HOUSEHOLD HEAD			
Older child (7-18)	0.1	0.1	<0.1
Male (18-59)	64.8	64.6	64.5
Female (18-54)	10.8	10.1	10.0
Male (60+)	11.6	12.0	12.3
Female (55+)	12.7	13.2	13.1
HOUSEHOLD SIZE			
1	17.5	17.0	16.0
2	26.9	27.2	27.8
3	23.8	23.1	22.9
4	21.0	21.4	21.5
5	7.0	7.2	7.6
6+	3.8	4.1	4.2
NUMBER OF CHILDREN <7			
0	78.5	78.8	78.5
1	17.8	17.5	17.7
2+	3.7	3.7	3.8
NUMBER OF CHILDREN 7-18			
0	65.2	64.6	64.1
1	22.4	22.5	22.6
2+	12.4	12.9	13.3
NUMBER OF WORKING-AGE MALES			
0	35.2	35.5	35.5
1	55.0	54.3	54.3
2+	9.8	10.2	10.2
NUMBER OF WORKING-AGE FEMALES			
0	34.7	35.5	35.6
1	56.4	55.5	55.6
2+	8.9	9.0	8.8

* Including households followed to new residences.

Tables 4 and 5 give a partial answer to the question. The second column in each shows a multinomial distribution or median value of a characteristic as measured for the Round V sample of cooperating households. The third column gives the same statistic (again the Round V characteristic) but computed only for households that participated in both Rounds V and VI. The final column gives the statistic based on Round V measures only for households that participated in all three rounds.

Here, as was the case for cross-sectional analysis, the most notable effect of attrition is the loss in the percentage of sample households from the Moscow/St. Petersburg region and the broader urban domain. Between Rounds V and VII there is also a modest loss in the relative percentage of single-person households. Round V-VII attrition does not appear to seriously distort the relative distribution of households by count of children or numbers of working men and women.

Table 5 shows the impact of Round V-VII attrition on the financial characteristics of the household "panel." It suggests that households that move out of their original residences or decline to participate at Round VI, or Rounds VI and VII, have higher median incomes and expenditures than households that remain in their original residences and continue to cooperate in the RLMS.

Table 5: Attrition in the Round V Panel, Round V Income Statistics for Respondents and Nonrespondents at Later Rounds

	Round V Panel	Round VI Panel		Round VII Panel	
Statistic	R	R	NR	R	NR
Round V Median Household Income	354,564	349,000	396,490	344,000	395,095
Round V Median Household Expenditure	466,593	465,552	474,404	463,657	498,451
Round V Median Income, % Poverty	2.024	1.995	2.179	1.976	2.138

Table 6 repeats the Table 4 analysis for a "panel" of individual respondents. As with households, nonresponse and movement have the greatest impact on the percent of individuals from the Moscow/St. Petersburg region and the more general urban domain. Attrition effects on the relative age/sex distribution produce a general aging of the "panel" of individuals. Consistent with the finding for households, nonresponse and movement result in losses of "panel" members from the higher economic ranks. Interestingly, there is only a slight disproportionate tendency for individuals who are unemployed at Round V to leave the sample at Round VI or VII. Those who remain at Rounds VI and VII are slightly older and are more likely to have had a normal body weight at Round V than are those who left after Round V.

Table 6: Attrition Effects for the Round V Individual Panel

	Percent by Category		
	Round V	Round VI	Round VII
Subpopulation	Panel	Panel	Panel
REGION			
Moscow/St. Petersburg	10.5	8.0	7.0
North/Northwestern	7.2	7.5	7.1
Central/Central Black Earth	18.1	18.4	19.1
Volga/Viask/Volga Basin	17.0	18.3	19.0
North Caususes	13.4	13.5	13.5

Table 6: (continued)

Subpopulation	Percent by Category Round V Panel		Round VI Panel		Round VII Panel	
Urals	14.4		15.0		15.5	
Western Siberia	9.9		9.8		9.9	
Eastern Siberia/Far East	9.6		9.5		9.0	
SETTLEMENT TYPE						
Urban	69.3		66.0		64.4	
PTG	5.5		5.8		5.9	
Rural	25.2		28.2		29.7	
AGE GROUP/SEX	M	F	M	F	M	F
0-19	14.5	14.0	14.2	14.3	14.2	14.2
20-39	13.9	15.6	13.0	15.0	12.4	15.0
40-59	11.1	13.6	11.2	14.2	11.2	14.8
60-79	5.5	9.5	5.7	10.2	5.6	10.6
80+	0.4	1.8	0.4	1.8	0.3	1.7
ECONOMIC RANK						
1	12.6		13.2		13.1	
2	15.4		15.8		15.8	
3	24.0		24.3		24.5	
4	22.5		21.6		21.6	
5	19.4		19.5		19.5	
6	4.1		3.9		3.8	
7+	1.9		1.6		1.6	
NORMAL WEIGHT?						
% Yes	54.5		56.4		57.4	
UNEMPLOYED?						
% No	96.2		96.5		96.7	
MEDIAN AGE	34		36		36	

3. REPLENISHING THE RLMS SAMPLE

As noted above, in the absence of housing construction, the original sample of dwelling units provides a cross-sectional representation of the Russian household population at each observed point. Of course, over a reasonable period there will be housing construction, and occupants of new units should be included in a sample that is to be nationally representative. Techniques such as those used in the U.S and Canada for sampling new housing construction could be employed to update the original sample of dwellings, but these techniques are complicated, and the necessary data (building permits, data from planning or housing agencies) may be difficult or expensive to collect today in Russia.

Most current housing construction in Russia is concentrated in multi-unit structures and development areas. It may be possible to replenish the sample by drawing a new sample of dwellings from the original enumeration lists compiled prior to Round V. New listings could be prepared for new housing structures located within the existing sample of second stage units (SSUs). A supplemental sample (at the correct rate for the SSU) could be selected from the new housing listing and combined with the sample from the listing of pre-existing housing to form an updated sample of dwellings.

Replenishment of the sample at some point may also be a good idea to avoid more serious problems of attrition among households that continue to reside in the original sample of dwelling units. The timing of replenishment will depend on several factors, not the least of which is cost.

4. WEIGHTS IN DESCRIPTIVE ANALYSIS OF RLMS DATA

Analysis weights are essential for unbiased sample-based estimation of RLMS descriptive statistics such as population and subclass means, proportions, and totals. The construction of a descriptive weight for cross-sectional analysis involves a simple sequence of steps: (1) determine the probability of selection for each sample household; (2) based on geographic and other known characteristics of sample households, compute an adjustment for nonresponding sample households; and (3) compute a nonresponse-adjusted weight as the product of the reciprocal of the sample selection probability and the nonresponse adjustment.

Since the RLMS attempts to interview all individuals within sample households, the selection probability for an individual equals that for his household. An individual in a cooperating household may, however, choose not to give an interview. If data on individuals--both cooperating and not--are known from household listings, the nonresponse adjustment factor in the analysis weight can be computed at the level of the individual. Fortunately, the majority of RLMS nonresponse at the individual level corresponds to noncooperation by the entire household, and the household nonresponse adjustment factor will capture most of the sample attrition loss at both levels.

If recent census data on households and individuals are available, a fourth post-stratification step can be added: scaling analysis weights so that the sum of weights for a defined subpopulation matches the corresponding census proportion (e.g., the weighted sample proportion of females, age 45 and older, in the Moscow/St. Petersburg region matches the corresponding proportion from the most recent census). The post-stratification of analysis weights serves two functions: (1) it can reduce the sampling variance of weighted estimates; more importantly, (2) it may correct noncoverage biases in the frame used to derive the original sample of dwellings and individuals.

There is considerable debate over the value of using weights in multivariate analysis. For example, in estimating linear or generalized linear models, many software programs allow the specification of weights for model fitting. Some statisticians argue that using weights is not necessary if the fixed effects that explain the variation in weights are included in the model. In RLMS data, the household characteristics that explain the greatest variation in weights are the geographic region and the urban/rural character of the civil division in which the dwelling is located. Variation in individual weights will reflect the geographic effects for households as well as differentials due to post-stratification of the sample by major geographic regions, age, and sex. Researchers who are interested in exploring the impact of RLMS weights on a multivariate analysis should consider the following test. Fit the model omitting the weights but including as fixed effects the household (region, urban/rural) or individual (region, urban/rural, age, and sex) characteristics. Without changing the specification, also estimate the model using the analysis weights. Compare the results to see if there are important differences in model parameters and/or interpretation. Differences in the unweighted and weighted versions could be due to added sampling variability introduced by the weighted estimation or could indicate that the model is not correctly specified.