Sample Attrition in the RLMS, 2001 - 2010.

Lessons for longitudinal analysis and an application in health.[1]

Christopher J Gerry [a,b,*] and Georgios Papadapoulos [c]

[a] Professor of Health Economics, National Research University – Higher School of Economics (St Petersburg), 16 Soiuza Pechatnikov, 190008 St Petersburg, Russia and [b] Deputy Director, Department of Economics of Health Reform, Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow, Russia.
[c] Lecturer in Econometrics, University of East Anglia, School of Economics, Norwich Research Park, Norwich, NR4 7TJ, UK
[*] Corresponding author. Tel.: +44 7525 926 952; Email: c.gerry@ucl.ac.uk

Abstract

The data of the Russian Longitudinal Monitoring Survey – Higher School of Economics represents one of the few nationally representative sources of household and individual data for Russia. It has been collected since 1992 and in recent years, thanks to more secure financial and logistical support, has become a resource increasingly drawn upon by scholars and students for national and cross-national studies. In this paper, we examine the extent of non-random attrition in the RLMS and discuss the circumstances under which this might give rise to biases in econometric analysis. We illustrate this with an example drawn from the health sphere.

Keywords: RLMS, Attrition Bias, longitudinal data, Health

JEL Classification Codes: C18, C23, C25, I12, J0, P30

[1] We are grateful to Carmen Li for the useful discussions we had when starting this work, as well as to Judith Shapiro and other participants at the first RLMS-HSE users conference at the Higher School of Economics, Moscow in May 2013, and to three anonymous referees.

1. Introduction

The Russian Longitudinal Monitoring Survey – Higher School of Economics (hereafter, RLMS) is a nationally representative series of comprehensive annual household surveys designed to monitor the health and economic welfare of households and individuals in Russia. It represents the only long-term nationally representative source of household and individual data for the Russian Federation and has become an important complement to equivalent longitudinal surveys from other countries. In recent years, thanks to more secure financial and logistical support, the RLMS data have become a resource increasingly drawn upon by scholars, students and practitioners both within and outside of Russia for national and cross-national studies, particularly in the fields of health, welfare, income and the labour market.

The richness of the data brings with it the opportunity to explore the causal processes that underlie the socioeconomic relationships observed during the period of so-called post-Communist 'transition'. However, the nature of this longitudinal survey also brings with it a level of complexity that demands attention and understanding by its users. At the heart of this concern, and the focus of this paper, is the problem of missing data due to attrition. All surveys are subject to missing data in the face of frequent non-response, but in the case of longitudinal data the issue is of particular importance as initially representative samples 'lose' respondents over time in a non-random manner. Such losses *may* induce sample selection bias due to attrition.

Consider a straightforward example. Imagine that we are interested in obtaining mean individual health outcomes, using the longitudinal element of the survey, with a view to tracking how respondents' health evolves over time. If the least healthy respondents leave the sample disproportionately in each year, due to so-called 'non-random attrition', then our mean health estimates will be biased because they will understate the proportion of poor health respondents. This will be true even if we are using the survey sample weights provided. Therefore, if we are interested in the longitudinal sample, we need to re-weight our estimates to account for the observed non-random attrition. For practitioners not interested in exploiting the longitudinal element of the data, the univariate statistics from individual cross-sections of the data could still be used without bias because of the annual replenishing undertaken to restore representativeness.

Of course, more commonly, we are concerned with regression based approaches intended to exploit the longitudinal element of the data to obtain consistent estimates of the coefficients of a conditional expectation (for example, mean health outcomes given gender, age, regional characteristics, time effects and so on). This is also problematic because analysis which ignores non-random attrition may produce estimates which are inconsistent and biased, if the model is not properly specified to account for the non-random attrition. Despite often being overlooked in empirical research, there is an important strand of literature modelling, and examining empirically, the effects of attrition in longitudinal survey analysis (Hausman and Wise, 1979; Nijman and Verbeek, 1992; Fitzgerald et al, 1998; Groves and Couper, 1998;

Lillard and Panis, 1998; Watson, 2003; Contayannis et al 2004; Behr et al 2005; Hawkes and Plewis, 2006; Jones et al, 2006). We discuss this further in sections 2 and 3 below.

In the RLMS data used in this paper, attrition can occur because the respondent has moved location, suffered family breakdown, died or is seriously ill, happens to be away and unavailable during the survey period, or has decided the survey is too costly in terms of time. This attrition can be permanent (absorbing state) or temporary, insofar as respondents may miss one or more rounds of the survey (perhaps due to temporary re-location, or short-term illness, or just unavailability at the time of the survey) before returning in subsequent survey rounds. Understanding the nature and consequences of this problem is crucial if longitudinal estimates that may unknowingly carry biases are to be avoided.

The main purpose of this paper is to conduct a detailed analysis of attrition and its determinants in the RLMS. In addition, we discuss the conditions under which common estimators for longitudinal data are inconsistent because of attrition. Specifically, we employ data from rounds 10 – 19 of the RLMS to: (i) systematically explore whether attrition in the RLMS is non-random in terms of socio-economic and demographic characteristics and; (ii) explore the potential effects of non-random attrition in an illustrative application relating to health. We are not aware of other studies that have rigorously examined attrition with the RLMS data for the post-2000 period. In view of the increasing accessibility and use of this survey, it is important that researchers and practitioners understand when attrition is changing the representativeness of the sample and which research questions are likely to be qualitatively affected by attrition.

We find strong evidence that attrition in the RLMS is systematically related to demographic, health, and other socioeconomic characteristics. We explain that whether this gives rise to biases in econometric work depends on the specific model under investigation and argue that, having a carefully specified model can minimise attrition bias. We illustrate this with respect to an example from health and find that, although attrition is non-random, the estimated effects of our regressors on health status are broadly robust across models, though not without some notes of caution. Our preliminary findings from the health application also offer support to the state dependence hypothesis and confirm the importance of unobserved individual heterogeneity.

We proceed as follows. In section 2 we introduce the RLMS survey and then examine general patterns of attrition in the data, before linking these descriptively to key socioeconomic and demographic characteristics. In section 3, we discuss the conditions under which the non-random attrition identified in section 2 may result in attrition bias, before outlining a methodological approach for testing and correcting for this bias. In section 4, we present and discuss the empirical implications of non-random attrition in the RLMS using an example from health. Section 5 concludes.

2. The RLMS Data

2.1 The longitudinal sample

The RLMS is a nationally representative series of comprehensive annual household surveys designed to monitor the health and economic welfare of households and individuals in Russia. Accordingly, each autumn, the survey collects rich information on a range of individual and household socio-economic, health and demographic variables. The survey strategy is predicated on the principle of 'repeated sampling of dwellings', in which all household members are interviewed in each survey (if they can be contacted within 3 visits), and then the dwelling itself (rather than the household) is followed. Combined with periodic (annual) replenishment this sampling strategy maintains the cross-sectional representativeness of the sample for each round. To further the longitudinal aims, there is a component of the panel which is followed regardless of dwelling and further attempts are also made to follow-up individuals who have moved out of the household.[2]

These somewhat complicated design features render the longitudinal element of the RLMS less straightforward than the most established household panel surveys (which typically follow the household rather than the dwelling)[3] and further complicate efforts to identify the nature of sample attrition. Compared to these other surveys, the RLMS data are *a priori* more likely to have high rates of attrition because of the dwelling oriented nature of the sampling strategy. It is also more likely to have substantial amounts of temporary attrition stemming from the follow-up efforts. We study attrition, among adults, for the years 2001 (round 10) to 2010 (round 19). We take round 10 as our starting point because the sample underwent a major replenishment at that time and it also represents the early stages of an extended period of consistent and regular annual surveys.

To identify our main longitudinal sample we take the full round 10 sample, replenished in order to ensure representativeness for that cross-section[4] and then, following it longitudinally we: exclude subsequent entrants into the sample, including those who reach adulthood after 2001; and include those that move out of the year-by-year representative sample after 2001 and that are followed within the RLMS. In what follows, when we refer to the '*longitudinal'* sample we are referring to the sample that is representative at round 10 (2001) and then followed, subject to attrition, through to round 19 (2010). In contrast, we refer to the annual cross-sectional survey data, which has been augmented to restore representativeness, as the

[2] The RLMS is a survey conducted by Higher School of Economics and ZAO Demoscope together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS (details and availability at http://www.cpc.unc.edu/projects/rlms-hse and http://www.hse.ru/org/hse/rlms).

[3] These include the British Household Panel Survey (now 'Understanding Society'), the US Panel Study of Income Dynamics, the Australian Household Income and Labour Dynamics survey, and the German Socio-economic Panel Study; respectively the BHPS, PSID, HILDA and SOEP. Note that the RLMS data are the latest addition to the Cross-National Equivalent File (CNEF) containing population panel data from Australia, Canada, Germany, Great Britain, Korea, Switzerland, and the United States. Further information is available at: http://www.human.cornell.edu/pam/research/centers-programs/german-panel/cnef.cfm.

[4] Note, in identifying the round 10 starting sample, we need to remove round 10 participants who had already moved out of the representative sample in previous rounds.

'*representative*' sample – the sample that includes survey replenishments and excludes participants that have moved out of the representative sample. The two samples are, by construction, the same in round 10, before diverging as the longitudinal sample becomes shaped by attrition.

This approach yields an initially representative longitudinal starting sample of 7,309 respondents, over the age of 17, in round 10 (59% of whom are female).[5] Taking this sample, we now examine the general patterns of attrition within the survey, before looking into the year-by-year descriptive statistics and comparing them to those obtained with the (cross-sectional) representative data.[6]

2.2 Patterns of attrition in the RLMS

Table 1 below summarises the basic patterns of attrition across various 'causes' (moving out of sample; death; split of household; unknown reason). This information is only available at any given round from household members that are still in the survey at the following round. That is, if someone dies or moves away after the previous survey period, then this gets recorded in the survey data the following period, if and only if, someone in the household reports it. Thus, if for any reason there are no remaining individuals in the household of a departed (through death or other means) person, or if the remaining individuals in the household did not want to reveal the departure, then this is recorded as 'unknown' attrition. The remaining attrition in the 'don't know' category could stem from tracing failure, failure to contact/follow-up or from survey non-cooperation.

Table 1 shows that round-by-round attrition is a little under 10% on average, which equates to overall attrition, after 9 years, of 49%.[7] It is clear that the cause of the overwhelming majority of attrition is formally unknown ('don't know'). This group will of course comprise many of those households that have either *all* moved or are all away. In these cases, with the whole household absent, there is no one present in the dwelling to provide information as to why they did not participate. This pattern is perhaps clearer still in figure 1 below. The left hand panel, showing the attrition rates between period t and period $t-1$, shows how the attrition hazard (just below 10% on average) declines over time, following an initial spike. This is to be expected: following the first round, or after a major replenishment, the least

[5] Our longitudinal sample includes 971 respondents that move permanently out of the year-by-year 'representative' data and that are followed within the RLMS, and a further 132 respondents who leave the representative sample, before subsequently returning.

[6] In any survey, there is an issue of whether respondents who choose to participate at the baseline are representative of the population but this is distinct from the issue of attrition, since in the latter case, at least the baseline characteristics of the non-responding group (attritors) are known.

[7] Placing this in the context of other longitudinal surveys, the figures are not out of kilter, particularly given the dwelling based sampling frame of the RLMS. For example, after 10 years of the BHPS survey, full interviews were carried out with a little over 60% of the original sample (Noah Uhrig, 2008; Jones et al, 2006). In the European Community Household Panel, from 2001 to 2008, dropout rates typically fall between 40 and 49%. Ireland had the highest (69%) and Portugal the lowest (30%). Moffitt et al (1999) using data from the PSID, found that 69% of the original (1968) sample were interviewed in 1978. In short, attrition is slightly higher in the RLMS but the difference is perhaps not as big as we might have anticipated.

'committed' respondents drop out first, leaving a more permanent survey base as time progresses. The right hand panel of figure 1 graphically captures the cumulative rate of attrition, confirming the extent of 'unknown' causes. The net effect of this attrition is that, after 9 years, the pooled (non-representative because of attrition and the inclusion of 'movers') longitudinal sample comprises of 50,181 adult observations (19,836 male / 30,345 female).[8]

Table 1: Attrition from representative sample

	Round 10	*Round 11*	*Round 12*	*Round 13*	*Round 14*	*Round 15*	*Round 16*	*Round 17*	*Round 18*	*Round 19*
No of Participants	7309	6260	5740	5315	4931	4657	4350	4012	3895	3715
No of Attritors		1049	1569	1994	2378	2652	2959	3297	3414	3594
Attrition rate		0.144	0.215	0.273	0.325	0.363	0.405	0.451	0.467	0.492
Moved		*0.019*	*0.027*	*0.037*	*0.043*	*0.046*	*0.051*	*0.059*	*0.059*	*0.064*
Died		*0.012*	*0.023*	*0.032*	*0.041*	*0.048*	*0.054*	*0.062*	*0.068*	*0.075*
Other/split		*0.005*	*0.009*	*0.008*	*0.007*	*0.006*	*0.006*	*0.007*	*0.008*	*0.008*
Don't know		*0.107*	*0.155*	*0.195*	*0.234*	*0.263*	*0.294*	*0.323*	*0.332*	*0.346*

[FIGURE 1 HERE]

A particular complication of the RLMS survey stems from the phenomenon of temporary attrition, whereby a respondent in the original longitudinal sample, returns to be surveyed, after having missed at least one previous interview. Indeed, while 40% of the initial longitudinal sample were 'always in' the sample and 43% left without returning during this period of analysis (so-called 'absorbing state' attrition), 11% were observed in the last round but had missed at least one round since first appearing, and a further 6% were not observed in the last round and had missed at least one round in between their first and last appearance. In terms of the pooled sample of 50,181, this means that 29,460 observations are from those always in the sample; 12,397 are from those that become permanent attritors; 6,104 are from temporary attritors who we observe in the last round; and 2,230 are from temporary attritors who we do not observe in the last round. Table 2 summarises the participation patterns (1 = participation; 0 = no participation) within the longitudinal sample and highlights the frequency of temporary drop-outs.

This feature of participation sequences in the RLMS is important. The bottom three rows, equating to more than 15% of the sample, all capture forms of temporary attrition. On the one hand, the scale of this type of attrition could be viewed positively, as it serves to limit the loss of the sample but, on the other hand, it raises the question of how to treat this category of respondents. For studies based on the PSID prior to 1990 (among others, Lillard and Panis, 1998; Fitzgerald et al, 1998) attrition was an absorbing state by construction, since households refusing the survey in one year, were not approached thereafter. Studies based on surveys which *have* followed up on 'refusing' respondents have tended to mirror this earlier

[8] For researchers interested in a shorter panel, round 15 represents a good option as the RLMS experienced another substantial replenishment at that time. The round 15 representative sample is 10,711 and the total longitudinal sample is 43,042. Space prohibits from discussing this further here, but we note that the correlates of attrition for the round 15 longitudinal panel are similar to those presented in this paper for the longer panel.

PSID based literature by making the assumption that the first non-response of a survey respondent signals permanent attrition, regardless of whether that respondent is subsequently contacted by the survey or not (Watson, 2003; Behr et al 2005; Hawkes and Plewis, 2006; Jones et al, 2006). This absorbing state assumption sidesteps the likely reality that temporary attritors have dropped from the sample for reasons that are quite different to those of 'permanent' attritors, not least because they are still alive.[9]

Table 2: Patterns of attrition in the RLMS

		Longitudinal Panel with Gaps (1)		*Longitudinal Compact Panel (2)*	
Label	*Pattern*	*Freq.*	*Percent.*	*Freq.*	*Percent.*
Always in	1111111111	2,946	40.29	2,946	40.29
First round only	1000000000	747	10.22	1049	14.35
First 2 rounds only	1100000000	460	6.29	688	9.41
First 3 rounds only	1110000000	409	5.6	590	8.07
First 4 rounds only	1111000000	325	4.45	485	6.64
First 5 rounds only	1111100000	249	3.41	357	4.88
First 6 rounds only	1111110000	270	3.69	368	5.03
First 7 rounds only	1111111000	254	3.47	367	5.02
First 8 rounds only	1111111100	166	2.27	211	2.90
First 9 rounds only	1111111110	248	3.39	248	3.39
1 round missing, there at end	1 – 0 – 1	399	5.46	-	-
>1 missing, there at end	1 – 00 – 1	370	5.06	-	-
Missing rounds, not there at end	1 – 0 – 1 – 0	466	6.38	-	-
Total		7,309	100	7,309	100

In the case of the RLMS, the researcher therefore needs to decide how to treat the respondents in the penultimate 3 rows of table 2. In this paper, in addition to identifying the longitudinal sample and the representative sample, we also identify the so-called 'compact' sample, where we treat all temporary attrition as absorbing state attrition. In other words, we drop all successive rounds for respondents that leave the survey, even though we know that they later return. In table 2, this results in 1,235 temporary attritors (bottom 3 rows) all being treated as 'left and never returned' attritors and thus distributed across the upper part of the table (that is, 302 are added to 'first round only', 228 to 'second round only', 181 to 'third round only' and so on). By dropping all future observations of these temporary attritors, we reduce the sample by more than 4,000 so, although it is common place to work with the compact panel, the researcher must be confident that doing so doesn't exacerbate any potential selection bias. We return to this discussion in section 4.

2.3 Descriptive statistics and attrition in the RLMS.

[9] In our case, the only respondents from among the 'left and never returned' category that can be considered as genuinely permanent attritors are those that died. That is, unless we have information concerning their death, in principle, the respondents referred to in lines 2-10 (first 2 – first 9 rounds) of table 2 could also be temporary attritors (in other words, they could return in the next round).

To get a better sense of attrition in the RLMS we compare the evolution of important socioeconomic and demographic characteristics for the longitudinal sample and the representative sample (with survey weights).[10] Appendix 1 presents these descriptive statistics for 2001 (round 10), 2004 (round 13), 2007 (round 16) and 2010 (round 19).[11] The variables for which we present statistics are typical of those derived from household survey data and include age, gender, settlement type, region, marital status, education level, occupational category, income, poverty status, unemployment status, household demographics, health and life-satisfaction indicators as well as variables referring to the respondents understanding and attentiveness during the survey.

Starting with the unweighted averages of the representative sample of round 10 we find the average age is 46.6 years and the sample is in majority female (58.6%). Two-thirds of respondents live in urban areas, with 11.9% living in Moscow/St. Petersburg. Two-thirds of the sample is married, with around 15% being widowed or single and 8% being divorced. The sample, as one would expect for a post-communist country, is mostly well-educated, with just 17.3% having the most basic level of secondary education (8 years) and correspondingly just 6% being in unskilled occupations. A substantial 26% of respondents are 'out of the labour force', in addition to the 23% which are of retirement age. As of 2001, the incidence of poverty was a little over 20% and reported unemployment was 4.5%. The sample reports being rather dissatisfied with life (50% declare less than average life satisfaction) and unhealthy, with incidences of chronic disease, high blood pressure and health problems (in the last 30 days) approaching 50%, though with less than 20% self-assessing their health as poor or very poor. Though expected (World Bank, 2013), these figures on life satisfaction and health, are still striking.

There are a few key differences between the unweighted and weighted 2001 means. Specifically, the unweighted sample is less male, older and more likely to be widowed, while less likely to be married or single; it is also less educated, with lower incomes, as well as being less healthy – both via objective and subjective measures. The inferior health status reflects that the sample is older, poorer and with lower levels of human capital. Therefore, if interest lies in cross-section univariate statistics, the survey weights should be used.

In terms of the evolution of the longitudinal sample over time, by design, the sample becomes older and this in turn impacts the composition of the sample. Compared to the (weighted) representative samples over time, the longitudinal sample becomes: older (and therefore also more retired and with lower numbers of children); less male; less urban (and correspondingly more rural); less likely to reside in Moscow/St. Petersburg (and more likely to reside in the Volga region, the North Caucasus and the Urals); less married and less single, while more

[10] That is, we use the survey weights, provided with the RLMS data, that correct for the unequal probability of dwelling selection based on population characteristics (gender, age distribution) known from census data.

[11] Sampling weights are not available for longitudinal purposes after 2001 because we retain people in the longitudinal sample that move dwelling but that are followed by the RLMS, and therefore have a sampling weight of zero (see footnote 5).

widowed and divorced; less engaged with the labour market and therefore less likely to be unemployed; and with declining relative health outcomes.

Figure 2, plotting the participation rates over time, presents a visual aid to understanding some of the emerging patterns that are described above. These graphs, and our subsequent analysis, are conditional on round 10 (initial) characteristics because these are the only characteristics that we observe for all participants (that is, we don't observe attritors once they attrit). For time invariant and highly persistent variables this is not controversial, but for variables that may change over time (including income; marital status; occupation) this is a less innocent strategy.

[FIGURE 2 HERE]

Figure 2(a), highlights the specificity of the Moscow/St. Petersburg regions – both of which are subject to very rapid attrition. The population in these areas is likely to be more mobile. Accordingly, figure 2(b) shows how urban respondents attrit more quickly than non-urban respondents. Figure 2(c) shows how it is the youngest and oldest age groups that leave the sample most quickly. The latter group likely captures the ailing health of the elderly, while the younger attrition may reflect the greater mobility of that population sub-group, or the difficulty in following up recent young home-leavers. Turning to our interest in health related attrition, figure 2(g) provides a clear visual hint that attrition is health related: respondents starting out with poor health in round 10 leave the survey more rapidly than healthier individuals do. To a lesser extent the same is observed for figure 2(h) and 2(i) which respectively detail attrition among those reporting recent health problems (in the last 30 days) and high blood pressure (ever told by doctor that they have high blood pressure). The attrition difference is clear for those reporting recent health problems but is marginal in the case of high blood pressure (though this disguises the higher attrition rates of males reporting that they have had high blood pressure). Figure 2(d) confirms that the longitudinal sample becomes less male over time. Figure 2(e) shows how respondents that were widows, single or divorced, attrit more rapidly in comparison to those whom were married at round 10. Figure 2(f) demonstrates that the least educated group also leave the survey disproportionately compared with other education groups. Figure 2(j) suggests that those most satisfied with life leave the survey at a higher rate while, consistent with this, figure 2(k) shows that it is the high income quintile that has the highest attrition rate. Finally, figure 2(l) shows that the unemployed attrit more rapidly than the employed.

In appendix 2 we combine the patterns of attrition, discussed in 2.2, with the socioeconomic and demographic characteristics introduced above through presenting the round 10 means conditional on the type of attrition observed. This largely confirms the findings of figure 2 but does add one or two interesting insights as to the differences between permanent and temporary attritors. Firstly, with regard to the age category data, the elderly ('Age>=60' and 'Retirement Age') are the dominant groups among those leaving and never returning, while the young are over-represented among temporary attritors. Secondly, the widowed are less prevalent among the 'always in' category, while married respondents account for a higher proportion of those 'always in' and a lower proportion of permanent attritors. Single and

divorced respondents are more heavily represented among temporary attritors. Thirdly, the least educated group make up a larger proportion of the permanent leavers than their sample presence predicts, while the unemployed, those in poverty, and the unskilled are more likely to be temporary attritors. Correspondingly, those in the highest income quintiles make up a very low proportion of those always in the sample. Indeed, 83.5% of those always in the sample are from the bottom 3 income quintiles. Finally, the unhealthy are more likely to permanently leave the sample, than to be temporary attritors.[12]

From this detailed examination of the nature and distribution of attrition within our longitudinal sample, we cull the following stylised descriptive facts: (i) males, the elderly, the least educated, those living alone and in poor health are the most likely respondents to leave the longitudinal sample without returning; (ii) the young, the single, those in urban areas and Moscow/St. Petersburg, those with university education and those in the top two income quintiles are more likely to be attritors in general, and *temporary* attritors in particular. We now go on to examine the correlates of attrition in a multiple regression framework.

2.4 Attrition in a multiple regression framework

The bivariate associations discussed above have confirmed our priors and those of the attrition literature (for a good example, see Groves and Couper (1998) for a thorough survey of demographic associations with attrition) that there are non-random patterns of attrition in the longitudinal sample. However, we need to go beyond the bivariate analysis above, and therefore estimate a series of (Probit) regressions in which we control for important demographic and socio-economic factors.

Specifically, we present round-by-round Probit 'participation' equations where the dependent variable is a dummy variable that takes value 1 if someone participates in round t and 0 otherwise, where $t = 11, 12, \ldots, 19$. We present the partial effects (evaluated at the means of the explanatory variables of round 10) from estimating these equations separately for males and females (because of the very different attrition patterns) in Appendix 3. The presented results are the 'fullest' of a number of specifications that we looked at and the findings discussed below are robust to more parsimonious specifications (for which fewer observations are dropped).

This regression framework analysis reinforces the bivariate findings discussed above. For both females and males: participation is more likely outside of the Moscow and St. Petersburg regions, in non-urban settlements more generally, for married individuals, for those with higher levels of education, for those below the retirement age and engaged in the labour market, for the married, for those with lower incomes (and those above the poverty line), for those reporting less than full life satisfaction and for those in better health. For males, all ages below 60 are less likely to have participated in the survey, while for females only those under 30 are less likely to have participated. For females, having children is

[12] These findings are confirmed by a multinomial regression analysis. Results are available on request.

associated with higher participation, while for males with larger household sizes, participation is also more likely. Once again, these results reflect the relative immobility associated with these household characteristics.

With one eye cast towards the health example we present in section 4, the health results from appendix 3 merit further comment. The negative effect of poor self-assessed health is stronger for males and shows, for example, that compared to those in very good health, men/women in very poor health are 27/20 percent less likely to respond to the survey in round 12. By round 19, these probabilities increase to 35 and 29 percent respectively. We also control for more objective health indicators (recently reported problem and high blood pressure). These variables are closely correlated with self-reported health and, though the results are difficult to interpret, they suggest that for women, those in 'actual' poor health are more likely to stay in the sample, conditional on other regressors, while for men, there is a very weak indication that those with high blood pressure are more likely to attrit. Moreover, these results hold true even when we don't control for self-assessed health.

In sum, the bivariate descriptive analysis and the evidence obtained from the regression models lead to the conclusion that there is non-random attrition in the RLMS sample and that it relates very strongly to certain key characteristics: health, age, region, labour market status, marital status, income, family size; and less strongly to other important characteristics: such as, life satisfaction and occupational category. However, since non-random attrition does not guarantee that econometric estimates of key relationships will be biased, this prompts the more important question of whether and when it matters. To address this question, we now identify and discuss the conditions under which attrition can generate sample selection bias.

3. When does attrition matter?

Having established that attrition in the longitudinal sample is non-random, in this section we discuss the conditions under which sample attrition results in conventional estimators for longitudinal data producing inconsistent estimates, a problem widely known as attrition bias. This type of potential bias is closely related to the general case of 'sample selection bias', arising in situations where a sample is not drawn randomly from the population of interest (Heckman, 1976, 1979). Sample selection can stem from various survey mechanisms: respondents can self-select into a survey (for instance, with web-based surveys); survey data can be systematically missing (Little and Rubin, 2002), for example because respondents refuse to provide answers to some questions; or samples can become non-randomly selected when individuals decide to drop out of a longitudinal sample (as shown in section 2 above).

It is not always or automatically the case that sample selection affects the consistency of regression based estimates. In fact, sample selection bias arises when the selection mechanism in operation depends on unobserved characteristics that also affect the particular outcome variable of interest. Consider, as we will do in section 4, the case where we want to estimate the determinants of health status in a panel regression framework where we know attrition takes place and is non-random. In these circumstances, attrition bias arises when

individuals in our sample exhibit some unobserved characteristics affecting both the probability of participation in future periods as well as the health outcome. We briefly explore this example in section 4 but first, we present a more detailed, but by no means comprehensive, explanation by way of practical guidance to the practitioner.[13]

First, we need to situate the problem within a panel regression framework. Assume that we are interested in the conditional expectation $E(y_{it}|\mathbf{x}_{it})$, with $i = 1, \dots, n$ and $t = 1, \dots, T$, where y_{it} is the scalar dependent variable of interest, and $\mathbf{x}_{it}$ is a $K \times 1$ vector of covariates which includes a constant. Assuming that the model is linear in parameters we can write,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector $K \times 1$ of parameters to be estimated once a random sample from the population is obtained, and $v_{it} = u_i + \epsilon_{it}$ is the scalar composite error, where u_i is the unobserved individual effect that is constant over time with $E(u_i) = 0$, and ϵ_{it} includes all unobservables that vary over time and across individuals with $E(\epsilon_{it}) = 0$. In general, using a random sample from the population, consistent estimation of $\boldsymbol{\beta}$ follows under the strict exogeneity assumption,

$$E(v_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = E(v_{it}|\mathbf{x}_i) = \mathbf{0}, \quad \mathbf{x}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}]', \tag{2}$$

which states that the error at t is independent from $\mathbf{x}_{it}$ not only at t, but also at any other period $t - j$ and $t + j$ where $t \neq j$. This is a sufficient condition for non-correlation between both error terms and $\mathbf{x}_{it}$.[14]

Now assume that all randomly selected individuals participate in period $t = 1$ with probability 1, but that thereafter they may drop out of the sample at any subsequent period. So, let s_{it} take value 1 if the individual participates in the survey and 0 if they attrit. The characteristics y_{it} and $\mathbf{x}_{it}$ are then only observed when the individual participates in the survey. This raises the question: under what conditions will our subsequent estimates be inconsistent due to attrition bias?

3.1 Missing Completely at Random (MCAR)

First, consider the case where the participation mechanism is totally random, so that the decision to attrit is independent of both $\mathbf{x}_{it}$ and v_{it} in all periods. In this case, we can apply the usual estimators for panel data on the available observations of the outcome variable and the covariates ($s_{it}y_{it}$ and $s_{it}\mathbf{x}_{it}$, respectively) and obtain consistent estimates since,

$$E(v_{it}|\mathbf{x}_i, \mathbf{s}_i) = E(v_{it}|\mathbf{x}_i) = \mathbf{0}, \tag{3}$$

where $\mathbf{s}_i = [s_{i1}, s_{i2}, \dots, s_{iT}]'$ is the vector of participation dummies.[15] This amounts to stating that $\Pr(s_{it} = 1|y_{it}) = \Pr(s_{it} = 1)$ and, since it holds without conditioning on the covariates,

[13] For a formal and complete treatment refer to Wooldridge (2010, chapter 19).
[14] Note that pooled OLS requires contemporaneous independence, while for the Fixed Effect estimator we require only ϵ_{it} to be strictly independent from the regressors.
[15] The estimates will be less efficient than those based on balanced panels, because of the loss of information.

it allows the researcher to obtain consistent unconditional estimates of the explanatory variables of interest. This scenario, of completely exogenous selection, is the MCAR case (Little and Rubin, 2002).

3.2 Missing at Random (MAR)

However, as in the case of the RLMS data (section 2 above), it is often clear that attrition is related non-randomly to certain socioeconomic and demographic variables (for example, health status). This being so, we need to assume that participation is non-random and is captured by the latent variable model,

$$s_{it} = \begin{cases} 1 & \text{if} \quad U_{it} = \mathbf{x}'_{it}\boldsymbol{\gamma} + \eta_{it} > 0 \\ 0 & \text{otherwise} \end{cases}, \text{ with } E(\eta_{it}|\mathbf{x}_i) = 0, \tag{4}$$

so that individual i participates at period t if his/her utility (U_{it}) from doing so is greater than zero. This utility, and therefore the selection mechanism, depends on the set of exogenous covariates plus an error term (η_{it}). If this error is independent of v_{it}, conventional estimators on the observed data yield consistent estimates because condition (3) still holds. That is, conditional on the set of exogenous covariates $\mathbf{x}_i$, selection becomes as if random, and consistent conditional estimates can be obtained. This provides the so-called MAR (Little and Rubin, 2002), or 'selection on covariates' (Wooldridge, 2007) case. It essentially amounts, to stating that $\Pr(S_{it} = 1|\mathbf{x}_{it}, v_{it}) = \Pr(S_{it} = 1|\mathbf{x}_{it})$ which can be shown to be a sufficient condition for the conditional expectation assumption above.[16]

3.3 Selection on Unobservables

It may be the case that, beyond the observable variables in our data, there are common unobservables that affect both the dependent variable of interest *and* the participation mechanism. In this case, η_{it} in (4) is correlated with v_{it} in (1), so that $E(v_{it}|\mathbf{x}_i, \mathbf{s}_i) \neq 0$ since, even conditional on $\mathbf{x}_i$, there are unobservable variables concurrently influencing the dependent variable and the selection mechanism. In this case, conventional panel data estimators are generally inconsistent and alternative models are employed, based on the two-step Heckit estimator by Heckman (1976, 1979), which make assumptions concerning the joint distribution of v_{it} and η_{it} (Hausman and Wise, 1979; Nijman and Verbeek, 1992; Vella and Verbeek, 1999). [17]

These models though do not provide a straightforward solution. First, identification requires at least one exclusion restriction from (1). That is, there must be at least one variable, let us say q_{it}, which affects s_{it} but is independent of v_{it}. Such a variable is difficult to find since most characteristics that affect the decision to stay in or leave the survey (such as, poor health) are also likely to affect the outcome variable of interest (in our case, self-assessed

[16] The latter is however a stronger assumption that requires knowledge of the probability function of the selection mechanism rather than the conditional expectation only.

[17] Note that if the selection mechanism depends on u_i only, the fixed effect estimator for linear panel models on the selected sample is still consistent as it eliminates u_i.

health).[18] Second, these models tend to assume a linear specification model whereas, with survey data, it is often the case that the variables of interest are categorical, and require nonlinear models.

3.4 Selection on Observables and Inverse Probability Weighting (IPW)

In view of the constraints hinted at above, hereafter in this paper we restrict ourselves to the selection on observables scenario. This is because it is the case which has attracted recent attention in panel data applications (Fitzgerald at al, 1998; Moffitt et al, 1999; Contoyannis et al., 2004; Jones et al., 2006), largely because it gives rise to the IPW estimation method (Wooldridge 2007, 2010). IPW methods are easily applicable in nonlinear models too and if used cautiously, can provide consistent estimates. It is therefore attractive in the context of most contemporary household level longitudinal surveys, including to users of the RLMS. We demonstrate its use with the RLMS data in section 4.

Taking the selection on unobservables scenario as our point of departure, assume there is a vector of observables $\mathbf{z}_{it}$, which includes $\mathbf{x}_{it}$, and is observed both when $s_{it} = 1$ and $s_{it} = 0$, such that,

$$\Pr(s_{it} = 1 | v_{it}, \mathbf{x}_{it}, \mathbf{z}_{it}) = \Pr(s_{it} = 1 | \mathbf{z}_{it}), \quad \text{for } t > 1.^{19} \tag{5}$$

So, conditional on the observables, $\mathbf{z}_{it}$, selection becomes random. This is the case of selection on observables in which the properties of $\mathbf{z}_{it}$ are quite distinct from the properties of q_{it} in the selection on unobservables scenario outlined above. The key point is that, since respondents may attrit after period 1 but $\mathbf{z}_{it}$ must always be observed if we are to run regressions, $\mathbf{z}_{it}$ is replaced with $\mathbf{z}_{i1}$ so that the vector of observables includes only the period 1 (initial period) information on $\mathbf{z}_{it}$. This underpins equation (5) with a strong assumption. Essentially, (5) now says that the vector of first period observables ($\mathbf{z}_{i1}$) needs to be a sufficiently good predictor of s_{it} so that, conditional on it, the probability distribution of s_{it} does not depend on either the unobservables or the observed covariates of any other period.

Crucially, there is a distinction between $\mathbf{z}_{i1}$ and q_{it} which renders condition (5) plausible. That is, in contrast to q_{it}, $\mathbf{z}_{i1}$ should be endogenous in (1), so that it is correlated with v_{it}. So, $\mathbf{z}_{i1}$ can include all the first period values of our observed covariates and dependent variables, $\mathbf{x}_{i1}$ and y_{i1} respectively, but it can also include any other variable that is a good predictor of selection but that is endogenous in (1). For example, if we are interested in the socioeconomic and demographic determinants of self-assessed health, more objective measures of health, such as if the respondent had a health problem, a chronic condition or high blood pressure should be good predictors both of participation and of self-assessed health. However, we are interested in $E(y_{it}|\mathbf{x}_{it})$ and not in $E(y_{it}|\mathbf{x}_{it}, \mathbf{z}_{i1})$ and therefore we

[18] Variables that are exogenous to the respondents, such as information obtained from the interviewer may be more likely to satisfy exogeneity. For example, if there was a differential incentivising mechanism, such as vouchers for interviewees. Conditional on appropriate observables, this would affect the probability of dropping out of the survey in future periods, but not the outcome variable.

[19] At $t = 1$, all respondents participate (that is, $s_{it} = 1$) and therefore there is no attrition.

do not want to control for $\mathbf{z}_{i1}$ and doing so may distort the parameter estimates.[20] Indeed, estimates of (1) will be inconsistent by construction, since v_{it} is correlated with s_{it} because they both depend on $\mathbf{z}_{i1}$.

With these complications in mind, we obtain consistent estimates of $\boldsymbol{\beta}$, under (5), via the IPW estimator. In the first step, for all individuals participating at $t = 1$, we run regressions appropriate for binary data, such as a Probit, of s_{it} on $\mathbf{z}_{i1}$, for every $t > 1$. We then obtain the predicted probabilities of participation at these periods for each individual, $\hat{p}_{it}$, and construct the inverse of the predicted probability of participation in period t as $\widehat{w}_{it} = 1/\hat{p}_{it}$. Finally, we then estimate a weighted regression model, in which the objective function is weighted by $\widehat{w}_{it}$, in much the same way as survey sampling weights might be applied to restore representativeness in a cross section. As explained above, we require that $\mathbf{z}_{i1}$ is endogenous in (1), because as Horowitz and Manski (1998) show, if $\mathbf{z}_{i1}$is exogenous, the IPW estimator reduces to the unweighted regression and therefore the more restrictive MAR condition must hold for consistency.

Finally, Fitzgerald at al. (1998) and Wooldridge (2010) also discuss a case where at any period t, $\mathbf{z}_{it}$ is constructed by using the available information at $t - 1$ and not just at $t = 1$. For this we revert to the restrictive assumption that attrition is an absorbing state, because otherwise there will be temporary attritors in the sample for whom no $t - 1$ observations are available. In this case (which we will refer to as IPW2) to construct our weights, for every period $t > 1$ we run a Probit regression of s_{it} on the information for $\mathbf{z}_{it}$ at $t - 1$ and predict a weight for round t as before; call it $\hat{\pi}_{it}$ for which $\pi_{it} = \Pr(s_{it} = 1|\mathbf{z}_{it}, s_{it-1} = 1)$. Then the weights to be used in the IPW2 are constructed sequentially. For example, for period 5, $\hat{p}_{i5} = \hat{\pi}_{i2} \times \hat{\pi}_{i3} \times \hat{\pi}_{i4} \times \hat{\pi}_{i5}$. However, we now require a form of strict exogeneity as (5) is transformed to $\Pr(s_{it} = 1|\mathbf{y}_i, \mathbf{z}_i, s_{it-1} = 1) = \Pr(s_{it} = 1|\mathbf{z}_{it}, s_{it-1} = 1)$, where $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{iT}]'$ and $\mathbf{z}_i = [\mathbf{z}_{i1}, \mathbf{z}_{i2}, \dots, \mathbf{z}_{iT}]'$, with $\mathbf{z}_i$ including $\mathbf{x}_i$. On the one hand therefore, this sequential construction may provide predicted probabilities that are stronger predictors of attrition, but on the other hand, a stronger assumption of strict exogeneity must hold for consistency.

3.5 Testing for Attrition Bias

Finally, because they are widely used, it is worth our noting two of the available tests that assess attrition bias. The most common and straightforward to apply of these was suggested by Verbeek and Nijman (1992). As assumption (3) requires that, given $\mathbf{x}_i$, $\mathbf{s}_i$ is independent of v_{it}, it is reasonable to assume that past or future values of s_{it}, or other functions of selection such as the total number of rounds participating in the survey, should not have any effect on y_{it} in (1). This is easy to apply but Verbeek and Nijman (1992) themselves warn us

[20] Moffitt at al. (1999) give a good example of this using the private returns to schooling in a Mincerian equation, where they do not want to include a variable for occupation even though it is arguably a good predictor of wage, because doing so will distort the causal interpretation of the effect of schooling, since one channel through which schooling affects wages is occupation type.

of the relatively low power of these tests.[21] Verbeek and Nijman (1992) also suggest Hausman-type tests between the estimates obtained from the longitudinal sub-sample and the estimates obtained from a balanced sample with all attritors removed. Again though, the authors themselves recognise that these tests have low power in the case that the asymptotic bias of both estimators is in the same direction.

4. An application to a model of Self-Assessed Health

Ultimately, whether or not attrition matters depends on the particular research question we are facing. If we find that attrition is non-random in our particular area of interest then we need to consider whether we can plausibly argue that the attrition is not correlated with the error term in equation (1), conditional on the observed variables. If we cannot make that case then we need to consider whether there are (endogenous) variables in our data and relevant to our question that can be included in the vector of observable characteristics ($\mathbf{z}_{i1}$).

When it comes to implementing the specific application, there are some important decisions to take regarding the sample itself. First, the sample should be restricted to those observations for which the dependent variable of interest has no missing cases (for reasons other than attrition) or alternatively, decide whether other interpolation methods for missing data are appropriate. Second, almost inevitably, there will be missing cases among the explanatory variables too which need addressing, otherwise a missing value for one variable in one period only will drop the entire observation from the sample, and the respondent will appear incorrectly as a temporary attritor.[22] Once the sample is identified, the Verbeek and Nijman (1992) tests offer a means of assessing possible attrition bias but, regardless of the results of their tests, the IPW estimators should still be applied.

In section 2 we described how attrition in the RLMS is related systematically to a number of themes, including health, region of residence, age, gender, labour market and education. The evidence suggests to us that attrition is a particular problem in the health sphere. For men, the effects of poor health and very poor health, on attrition, are statistically significant (relative to the reference category of very good health) across the entire period, while for women, the effect of very poor self-assessed health is significant from round 11 (2002) onwards. In both cases, the impact of poor initial health increases as the panel lengthens (Appendix 3). With these results in mind, we draw on this health example to explore the implementation of the selection on observables case.

4.2 Self-assessed health in Russia[23]

[21] Moreover, note that adding these indicators in (1) does not solve the attrition problem since these variables should not be seen as representing the exact dependence mechanism between v_{it} and s_{it}. If that had been known, then we could have applied a two-step Heckit procedure.

[22] In our case, we face only very small amounts (less than 200 cases in total) of this type of 'attrition'.

[23] In the application that follows, for the sake of brevity in our illustration, we omit important econometric and methodological points that the researcher should be aware of. We refer the reader to fuller coverage in Contoyannis et al. (2004) and Wooldridge (2010).

The deteriorations and fluctuations in health outcomes in the post-Communist world have been well documented and debated (Leon et al. 1997; Cornia and Paniccià 2000) but few Russian specific studies have emerged in the economics literature based on the RLMS data. Lokshin and Ravallion (2008) develop an estimation method that allows them to argue that, despite evidence to the contrary in the raw self-assessed health data, there is robust evidence of an economic gradient in health status in Russia. Denisova (2010) is possibly the first work to examine the determinants of Russian mortality controlling for both individual and household heterogeneity. She finds that relative status and the associated chronic stress, unemployment and immobility in the labour market and the excessive use of alcohol and smoking are the key determinants of mortality in the RLMS data. Denisova includes a brief discussion of attrition, acknowledging its potential importance, while surmising that its impact is likely to be low on the basis that there is no significant health difference between those leaving the sample and those in the total sample.

It is not the purpose of this illustration to add to those comprehensive analyses of health determinants, but rather illustrate the lessons of non-random attrition for longitudinal studies, within the context of self-assessed health (SAH) determinants in Russia. To model SAH, for illustrative purposes, we first transform the 5-category variable into a dummy variable, (H_{it}), that takes the value of 1 if the individual reports poor or very poor health ('unhealthy'), and 0 otherwise.[24] More formally, we firstly specify H_{it}^{*} as the latent self-assessed health variable which, adapting equation (1) from section 3, is given as follows:

$$H_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + H_{it-1}\gamma + u_i + \epsilon_{it}, \text{ where } (i = 1, \dots, N; t = 2, \dots, \mathrm{T}). \quad (6)$$

Here, $\boldsymbol{\beta}$ is the $K \times 1$ vector of parameters to be estimated, ϵ_{it} is assumed to be normally distributed with mean zero and strictly exogenous with respect to the explanatory variables, and H_{it-1} is the dummy indicator for lagged health. We include the lagged health term because we know that health status is likely to be persistent over time (Gerry, 2012) either due to pure state dependence or unobserved heterogeneity. Table 3 below, which shows the transition from the one state of health to the other, confirms the very high persistence of the dependent variable. Although untangling these two causes is beyond the scope of this paper, equation (6) allows us to separately identify state dependence (measured by the effect of H_{it-1}) through the inclusion of unobserved heterogeneity captured in the individual effect, u_i.

Table 3: Transition between health states

	Not Unhealthy (t)	Unhealthy (t)
Not Unhealthy (t-1)	**92.25%**	7.75%
Unhealthy (t-1)	29.89%	**70.11%**
Total	80.88%	19.12%

[24] Our estimates are based on the sample for which we have complete information on health status. This yields a full sample of 50,119.

From the latent model of equation (6), we have that $H_{it} = 1$ if $H_{it}^{*} > 0$ and zero otherwise. Therefore,

$$\Pr(H_{it} = 1) = \Pr(H_{it}^{*} > 0) = \Pr(\epsilon_{it} > -\mathbf{x}_{it}'\boldsymbol{\beta} - H_{it-1}\gamma - u_i), \quad (7)$$

which, assuming ϵ_{it} is standard normally distributed, yields a Probit model with $\Pr(H_{it} = 1) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + H_{t-1}\gamma + u_i)$. At this point we need to deal with u_i, since it is not independent of past health status and is potentially correlated with the remaining variables. Furthermore, in a dynamic specification, an additional problem arises in the form of so-called 'initial conditions' (Heckman, 1981). Wooldridge (2005) proposes a 'simple approach' to dealing with the above problems, which allows for the possibility that the individual effect is correlated with the observed explanatory variables as well as with the lagged effect. Following the approach of Mundlak (1978), Wooldridge (2005) specifies $\mathbf{w_i}$ as the leads or lags of the exogenous explanatory variables. The correlation between the initial observation H_{i1} and α_i is allowed in order to render an error term which is uncorrelated with H_{i1}. Here, instead of including $\mathbf{w_i}$, following Contoyannis at al. (2004), we assume that $u_i = \alpha_0 + \alpha_1 H_{i1} + \bar{\mathbf{x}}_i'\boldsymbol{\alpha}_2 + c_i$, where $\bar{\mathbf{x}}_i$ is a vector of the averages of the time variant variables. Thus, the model has a correlated random effect structure in which the regressors at time *t* include the initial conditions as well as the $\bar{\mathbf{x}}_i$ vector:

$$H_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + H_{it-1}\gamma + \alpha_0 + \alpha_1 H_{i1} + \bar{\mathbf{x}}_i'\boldsymbol{\alpha}_2 + c_i + \epsilon_{it} \quad (8)$$

Therefore, the Probit model, which we refer to as the correlated effects Mundlak-Wooldridge model, now becomes $\Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + H_{it-1}\gamma + \alpha_0 + \alpha_1 H_{i1} + \bar{\mathbf{x}}_i'\boldsymbol{\alpha}_2 + c_i)$.[25] If we treat c_i as a RE we obtain a RE Probit, while if we simply ignore it, we arrive at the, less efficient, pooled Probit. This would be fine in the case of static estimates. However, if the model is dynamic, then pooled OLS will not provide consistent estimates, unless c_i is zero. Resolving this problem is beyond the scope of this illustrative example, so we follow the approach of Contoyannis et al. (2004) and employ dynamic longitudinal pooled Probit and RE Probit specifications on both the longitudinal sample and the compact sample and seek to understand whether the systematic patterns of non-random health-related attrition detailed in section 2, result in attrition bias. In estimating the IPW models we obtain standard errors clustered by individuals in order to allow for serial correlation.

In implementing (8), our specifications are appropriately parsimonious, including alongside the lagged dependent variable and its initial condition, the time invariant variables relating to gender, region and education as well as the continuous age variable and the round dummies.[26] Before turning to the estimates, we start by conducting the very simple 'variable addition' tests referred to in section 3 (Verbeek and Nijman, 1992). These test for the existence of a relationship between an individual's health status and two indicators of attrition: whether they

[25] Note that because of including α_0 we need to restrict the constant from vector $\boldsymbol{\beta}$ to zero. In addition, $\bar{\mathbf{x}}_i$ cannot include round dummies because that would cause colinearity.

[26] We also tried specifications with quadratic and cubic terms for age, but they provided a poor fit while another specification with dummies for age, suggested that the relationship between health and age is linear.

responded to the survey in all rounds; and the sum of rounds in which they are present. In essence, we create indicator variables for the latter and then add these variables (separately) to dynamic Probit model specifications that include correlated effects and initial conditions. The results are presented in table 4 and are all suggestive of attrition 'bias', though tell us little about the extent or nature of that bias.

Table 4: Tests for attrition

Verbeek & Nijman Attrition Test	Dynamic Pooled Probit			Dynamic RE Probit		
	Coeff	Rob.SE	P-value	Coeff	Rob.SE	P-value
=1 if Always In, =0 otherwise	-0.113	0.030	0.000	-0.205	0.048	0.000
Number of Rounds Participated	-0.045	0.013	0.001	-0.073	0.020	0.000

To further explore the influence of attrition, table 5 presents the coefficient estimates for Probit models based on pooled and RE specifications. For the pooled Probits we present the IPW and IPW2 specifications.[27] The IPW estimates use round 10 regressors to predict attrition, while IPW2 also includes values from the previous round.[28] We estimate the pooled and weighted pooled specifications on both the longitudinal and the compact panel. Further to the discussion in section 3, to construct the weights we include additional variables that are good predictors of participation but also potentially endogenous in equation (8): dummies for all categories of health status, a dummy for whether the respondent had a health problem in the last 30 days, gender, age, regional dummies, settlement type, marital and occupational status, education, family size and a dummy taking value one if the respondent reports having children.[29]

First and foremost, table 5 suggests that the non-random health related attrition detailed above may not bias the estimates of models that do not adjust for attrition, since the unweighted and weighted estimates are similar.[30] Further, comparing the estimates of the longitudinal and compact sample Probits also suggests that making the 'absorbing state' assumption concerning the nature of attrition does not substantially affect the results, although the IPW2 estimates show reduced significance when using the compact panel. Notwithstanding this, there are one or two important distinctions between the weighted and unweighted estimates which merit mention. Being male, which is associated with a lower

[27] Note that, due to the different scaling of the error variance, the estimated coefficients of the RE model and the pooled model are not directly comparable so we simply compare the relative effects of pairs of variables across the two models.

[28] Strictly speaking, the standard errors of both IPW models must be adjusted for the fact that we have used the predicted probabilities to construct the weights, rather than the true ones. However, interestingly, Wooldridge shows that the model with the predicted weights is more efficient than if we had known the true weights. Therefore the unadjusted standard errors can actually be considered as 'conservative' ones.

[29] We experimented with weights derived from using extra variables (a poverty indicator, a proxy for high blood pressure, life satisfaction, income quintiles, and whether the interviewer thought that respondents had a good attitude or understanding in the interview), but do not present them here because the inclusion of these variables increases the number of missing observations. We therefore simply note that the results are qualitatively similar.

[30] Note that, this does not imply that our estimates are consistent, since there might be other forms of misspecification, but it does imply that attrition may not exacerbate any potential inconsistencies. However, the seeming absence of attrition could simply reflect that the bias operates in the same direction so as to be invisible in the IPW estimates.

likelihood of being (self-assessed) unhealthy is not significant in the IPW-2 models, [31] while the regional effects are substantially different, as the regions become less significant (relative to Moscow) in the weighted model, and in IPW2, the signs also change. IPW2 also differs in not confirming a temporal improvement in health. We should therefore interpret the results relating to time, region and gender effects with due caution.[32]

Turning now, briefly, to the other results of these estimates, there are a number of important findings, robust to all of the different approaches and specifications. Controlling for state dependence is important as, in all of our models, the lagged health variable has very strong effects on current health in all models. The initial conditions (health in round 10) are also significant across the models, which more than likely picks up some of the effect of unobserved heterogeneity. Interestingly, in the RE Probit these initial health conditions are in fact more important than lagged health. In other words, even controlling for unobserved heterogeneity, through the Mundlak-Wooldridge approach, there remains strong evidence of state-dependence in health. This mirrors the findings of Contoyannis et al. (2004) for the British household panel data. Age increases the likelihood of reporting bad health, though the average age coefficient is negative, suggesting that the unobserved heterogeneity related to age is important. We also confirm that having the lowest level of education or the low-skilled category of vocational education is associated with low SAH.

Finally, as noted above, in the RE Probit model, the estimates of ρ are highly significant, suggesting that, while the inclusion of the mean of the age variable and the initial conditions capture something of the unobserved heterogeneity, individual heterogeneity remains important, accounting for approximately 46% of the latent error variance.

5. Conclusion

This paper was motivated by the increasing availability and use of longitudinal surveys, including most recently the RLMS. These surveys are inevitably subject to non-random attrition in their longitudinal elements, but understanding the nature and significance of that attrition is not straightforward and is often overlooked. In this paper, we examined attrition in the RLMS, discussed the scenarios when it is likely to matter and then considered techniques to statistically test and correct for it.

We found strong evidence that attrition in the RLMS is indeed non-random and in particular appears to be related to age, gender, health, education, marital status, labour market activity, region of residence and settlement type. Applying the inverse probability weighted (IPW)

[31] It is well known that males in Russia have considerably higher levels of ill-health than females, yet in self-reported surveys, this is rarely reflected.

[32] As mentioned previously, the RE estimates and the pooled estimates are not directly comparable and that to evaluate the magnitude of the associations between SAH and the regressors we would need to calculate the partial effects. Arulampalam (1999) proposes a transformation of the RE estimates to render comparability with the pooled estimates. Applying this, we find that the two sets of coefficients are similar, apart from the effect of previous health and the initial conditions. This difference is potentially important, though beyond the scope of this paper.

estimators, explained in section 3, we found that, although attrition is non-random, neither temporary nor permanent attrition bias the main health results of interest though they do raise a note of caution regarding the role of time, region and gender. Though we find that our weighted and unweighted estimates were similar we note that this does not automatically mean that there is no attrition bias; it could simply be that the bias operates in the same direction so as to be invisible in the IPW estimates.

With regard to our health estimates our headline findings are that: i/ as in Contoyannis et al. (2004) we find strong support for the state dependence hypothesis; ii/ there is a large role for unobserved individual heterogeneity, which may be associated with age, and with initial health status; iii/ as expected, and consistent with economic theory, age and low levels of education are associated with poor SAH.

Finally, the main aim of the paper was to draw attention to an overlooked phenomenon in longitudinal data analysis and, in the context of a dynamic health regression framework, to highlight a number of best practice procedures and techniques for practitioners using the RLMS. These include: first, taking due care in identifying the longitudinal sample; second, understanding the nature and variants of attrition within that sample; third, making careful decisions regarding how to treat missing observations that are not related to attrition; fourth, applying, but understanding the limitations of, regression framework models to examine whether non-random attrition is associated with attrition bias in the particular research question at hand.

Table 5: Dynamic Probit Models with Initial Conditions and Mundlak-Wooldridge Correlated Effects

	Unbalanced Sample with Gaps (or, Longitudinal Panel with Gaps)								Unbalanced Sample with Monotone Attrition (or, Longitudinal Compact Sample)			
	Pooled Probit		Weighted Pooled Probit - IPW1		Weighted Pooled Probit –IPW2		RE Probit		Pooled Probit		Weighted Pooled Probit –IPW2	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Unhealthy (*t-1*)	1.295***	(0.029)	1.341***	(0.032)	1.394***	(0.051)	0.560***	(0.030)	1.307***	(0.030)	1.370***	(0.041)
Unhealthy (*1*)	0.618***	(0.030)	0.630***	(0.032)	0.616***	(0.044)	1.389***	(0.052)	0.622***	(0.031)	0.657***	(0.040)
Age	0.100***	(0.011)	0.104***	(0.012)	0.092***	(0.014)	0.149***	(0.016)	0.095***	(0.011)	0.094***	(0.014)
Male	-0.099***	(0.025)	-0.051*	(0.027)	0.009	(0.049)	-0.151***	(0.038)	-0.113***	(0.026)	-0.027	(0.044)
Northern & North Western	0.178***	(0.060)	0.098	(0.071)	-0.077	(0.107)	0.275***	(0.095)	0.278***	(0.067)	0.067	(0.108)
Central & Central Black-Earth	0.222***	(0.046)	0.147***	(0.054)	-0.056	(0.083)	0.319***	(0.075)	0.320***	(0.053)	0.068	(0.096)
Volga-Vaytski & Volga Basin	0.236***	(0.047)	0.142***	(0.054)	-0.088	(0.083)	0.327***	(0.075)	0.337***	(0.053)	0.036	(0.107)
North Caucasian	0.080	(0.051)	-0.022	(0.057)	-0.223**	(0.087)	0.117	(0.079)	0.156***	(0.058)	-0.124	(0.110)
Ural	0.095*	(0.049)	-0.005	(0.057)	-0.224***	(0.085)	0.141*	(0.079)	0.191***	(0.056)	-0.093	(0.107)
Western Siberian	0.199***	(0.055)	0.072	(0.063)	-0.179**	(0.089)	0.278***	(0.089)	0.304***	(0.061)	-0.032	(0.111)
Eastern Siberian & Far Eastern	0.204***	(0.053)	0.105*	(0.062)	-0.085	(0.097)	0.325***	(0.086)	0.312***	(0.060)	0.023	(0.123)
Technical and Medical	0.037	(0.034)	0.076**	(0.037)	0.113***	(0.067)	0.053	(0.054)	0.020	(0.036)	0.016	(0.058)
Vocational – technical	0.130***	(0.039)	0.143***	(0.042)	0.181**	(0.077)	0.214***	(0.061)	0.116***	(0.042)	0.119*	(0.071)
Vocational – manual	0.188***	(0.056)	0.231***	(0.064)	0.371***	(0.154)	0.309***	(0.087)	0.179***	(0.058)	0.201**	(0.091)
High School	0.049	(0.043)	0.108**	(0.047)	0.164**	(0.077)	0.069	(0.065)	0.016	(0.046)	0.107	(0.077)
Incomplete high school	0.219***	(0.039)	0.279***	(0.042)	0.297***	(0.072)	0.370***	(0.061)	0.199***	(0.041)	0.233***	(0.066)
Mean of Age	-0.073***	(0.011)	-0.077***	(0.012)	-0.064***	(0.015)	-0.107***	(0.016)	-0.063*	(0.038)	-0.065***	(0.014)
Round 12	0.057	(0.038)	0.066*	(0.039)	0.078*	(0.041)	0.037	(0.041)	-0.127***	(0.038)	0.0798*	(0.040)
Round 13	-0.110***	(0.038)	-0.069*	(0.041)	-0.032	(0.044)	-0.169***	(0.049)	-0.173***	(0.043)	-0.048	(0.043)
Round 14	-0.196***	(0.044)	-0.163***	(0.046)	-0.107**	(0.050)	-0.299***	(0.059)	-0.200***	(0.050)	-0.063	(0.043)
Round 15	-0.213***	(0.051)	-0.156***	(0.054)	-0.093	(0.061)	-0.348***	(0.070)	-0.254***	(0.057)	-0.083*	(0.049)
Round 16	-0.267***	(0.059)	-0.179***	(0.063)	-0.020	(0.081)	-0.416***	(0.083)	-0.261***	(0.065)	-0.019	(0.082)
Round 17	-0.300***	(0.068)	-0.225***	(0.073)	-0.054	(0.093)	-0.457***	(0.097)	-0.402***	(0.074)	-0.076	(0.082)

Round 18	-0.435***	(0.077)	-0.370***	(0.082)	-0.209**	(0.101)	-0.612***	(0.111)	-0.531***	(0.085)	-0.225**	(0.094)
Round 19	-0.554***	(0.088)	-0.433***	(0.096)	-0.116	(0.140)	-0.785***	(0.126)	-0.069***	(0.011)	-0.167	(0.148)
Constant	-2.917***	(0.071)	-2.896***	(0.076)	-2.902***	(0.098)	-4.162***	(0.120)	-3.037***	(0.076)	-2.950***	(0.105)
Sample Size	42,643		42,483		42,483		42,643		38,550		38,416	
Log-Likelihood	-12,264.5		-20,342.2		-26,089.0		-11,605.6		-11,089.1		-20,582.6	
$\hat{\rho}$							0.457***					

Notes: For the Pooled models robust standard errors (clustered by respondent's ID) are presented in parentheses, calculated using the Delta method. * denotes 10% level of significance, ** denotes 5% level of significance, *** denotes 1% level of significance. The excluded categories for the dummy variables are: not unhealthy, female, Moscow & St. Petersburg, University, Round 11 (Round 10 is dropped because we include one lag of the dependent variable. Sample sizes change slightly because of missings in variables that are used to construct the weights. For monotone attrition, more are missing because we drop all subsequent rounds for temporary attritors when they drop out once. Results are robust to different specifications and sample sizes. We do also estimate RE Probit and IPW1 for the compact panel (estimates available on request) and the results are very similar to the ones obtained from their counterparts using the panel with gaps.

References

Arulampalam, W. (1999). 'Practitioners' Corner. A note on estimated coefficients in random effects Probit models', Oxford Bulletin of Economics and Statistics, 61(4), pp.597-602.

Behr, A., Bellgardt, E., and Rendtel, U. (2005). 'Extent and Determinants of Panel Attrition in the European Community Household Panel', European Sociological Review, 21, pp. 489-512.

Contoyannis P., Jones A.M., and Rice N. (2004). 'The dynamics of health in the British Household Panel Survey', Journal of Applied Econometrics, 19, pp. 473-503.

Cornia, G.A., and Paniccià, R. (eds.) (2000). The Mortality Crisis in Transitional Economies, Oxford, Oxford University Press.

Denisova, I. (2010). 'Adult Mortality in Russia: A Microanalysis', Economics of Transition, 19(2), pp. 333-363.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). 'An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics', The Journal of Human Resources, 33(2), pp. 251-299.

Gerry, C.J. (2012). 'The Journals are full of great studies but can we believe the statistics? Revisiting the Mass Privatisation – Mortality Debate', Social Science and Medicine. 75, pp. 14-22.

Groves, R. M. and Couper, M. P. (1998). Nonresponse in Household Interview Surveys, New York: John Wiley & Sons.

Hawkes, D. and Plewis, I. (2006). 'Modelling Nonresponse in the National Child Development Study', Journal of the Royal Statistical Society Series A, 169(3), pp. 479-491.

Hausman, J.A., and Wise, D.A. (1979). 'Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment', Econometrica, 47, pp. 455-474.

Heckman, J.J. (1976). 'The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Dimple Estimator for such Models', Annals of Economic and Social Measurement, 5(4), pp.475-492.

Heckman, J.J. (1979). 'Sample Selection Bias as a Specification Error', Econometrica, 47(1), pp. 153-161.

Heckman, J. (1981). 'The incidental parameters problem and the problem of initial conditions in estimating a time-discrete data stochastic process'. In Manski, C.F. & McFadden, D. (eds), Structural Analysis of Discrete Data with Econometric Applications, MIT Press: Cambridge, MA.

Horowitz, J. L., and Manski, C. F. (1998). 'Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation using Weights and Imputations', Journal of Econometrics, 84(1), pp. 37-58.

Jones, A.M., Koolman, X., and Rice, N. (2006). 'Health related non-response in the British Household Panel Survey and European Community Household Panel: using inverse-probability-weighted estimators in non-linear models', Journal of Royal Statistical Society Annals, Series A: Statistics in Society, 169(3), pp. 543-569.

Leon, D., Chenet, L., Shkolnikov, V., Zakharov, S., Shapiro, J., Rakhmanova, S., Vassin, S.,

and McKee, M. (1997). 'Huge variation in Russian mortality rates 1984-1994: Artefact, alcohol or what?', Lancet, 350, pp. 383-88.

Lillard, L.A., and Panis, C.W.A. (1998). 'Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality', Journal of Human Resources, 33(2), pp. 437-457.

Little, R.J.A., and Rubin D.B. (2002). Statistical analysis with missing data, John Wiley & Sons; New York.

Lokshin, M., and Ravallion, M. (2008). 'Testing for an Economic Gradient in health status using subjective data', Health Economics, 17(11), pp. 1237-1259.

Moffitt, R., Fitzgerald J., and Gottschalk, P. (1999). 'Sample attrition in panel data: the role of selection on observables', Annales D'Economie et de Statistique, 55-56, pp. 129-152.

Mundlak, Y. (1978). 'On the pooling of time series and cross-section data'. Econometrica, 46(1), pp. 69-85.

Nijman, T.E., and Verbeek, M.J.C.M. (1992). 'Non-response in panel data: The impact on estimates of a life cycle consumption function', Journal of Applied Econometrics, 7(3), pp. 243-257.

Noah Uhrig, S.C. (2008). 'The Nature and Causes of Attrition in the British Household Panel Survey', Institute for Social and Economic Research, Working Papers, No. 2008-05, Colchester: University of Essex.

Rendtel, U. (2002). 'Attrition in Household Panels: A Survey', CHINTEX (Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel – Implications on Data Quality), Financed by the European Commission under contract number IST-1999-11101, Working Paper No.4, Work Package 6. https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/WorkingPaper4.pdf?__blob=publicationFile

Vella, F., and Verbeek, M. (1999). 'Two-step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias', Journal of Econometrics, 90(2), pp. 239-263.

Verbeek, M., and Nijman, T. (1992). 'Testing for Selectivity Bias in Panel Data Models'. International Economic Review, 33(3), pp. 681-703.

Watson, D. (2003). 'Sample Attrition between Waves 1 and 5 in the European Community Household Panel', European Sociological Review, 19(4), pp. 361-378.

Watson, N., and Wooden, M. (2011). 'Re-engaging with Survey Non-respondents: The BHPS, SOEP and HILDA Survey Experience', SOEP papers No379. DIW, Berlin. http://www.diw.de/sixcms/detail.php?id=diw_01.c.373175.de

Wooldridge, J.M. (2005). 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', Journal of Applied Econometrics, 20, pp. 39-54.

Wooldridge, J.M. (2007). 'Inverse probability weighted estimation for general missing data Problems', Journal of Econometrics, 141(2), pp. 1281-1301.

Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data, The MIT Press, Cambridge and London.

World Bank, The. (2013). The Global Burden of Disease: Generating Evidence, Guiding Policy – Europe and Central Asia Regional Edition', Institute for Health Metrics and Evaluation, Human Development Network. Seattle, WA: IHME.

Appendix 1: Variable Names and Means by Round: Longitudinal (Long.) and Representative (Rep.) samples

	Round 10 Long.	Round 10 Rep.	Round 13 Long.	Round 13 Rep.	Round 16 Long.	Round 16 Rep.	Round 19 Long.	Round 19 Rep.
Male	0.414	0.450	0.395	0.451	0.388	0.454	0.385	0.452
Age	46.6	44.4	49.4	45.3	51.7	45.1	53.7	44.3
Urban	0.670	0.679	0.634	0.682	0.611	0.679	0.604	0.688
Rural	0.268	0.258	0.298	0.256	0.319	0.262	0.320	0.250
PGT (urban-type area)	0.062	0.063	0.068	0.062	0.070	0.059	0.076	0.059
Moscow & St. Petersburg	0.119	0.120	0.085	0.132	0.067	0.121	0.069	0.102
Northern & North Western	0.060	0.061	0.059	0.055	0.056	0.060	0.057	0.067
Central Black-Earth	0.182	0.180	0.185	0.171	0.181	0.181	0.182	0.199
Volga-Vaytski / Volga Basin	0.170	0.169	0.192	0.167	0.196	0.168	0.200	0.159
North Caucasian	0.141	0.139	0.153	0.140	0.165	0.152	0.167	0.149
Ural	0.138	0.141	0.152	0.138	0.156	0.136	0.153	0.145
Western Siberian	0.091	0.092	0.083	0.105	0.084	0.092	0.083	0.089
Eastern Siberian & Far East	0.098	0.098	0.091	0.093	0.095	0.090	0.089	0.092
Married	0.632	0.644	0.646	0.616	0.654	0.604	0.583	0.523
Single	0.147	0.158	0.109	0.176	0.093	0.196	0.093	0.238
Divorced	0.079	0.079	0.089	0.086	0.084	0.079	0.135	0.120
Widowed	0.142	0.118	0.156	0.126	0.169	0.121	0.188	0.119
University	0.224	0.230	0.233	0.236	0.224	0.245	0.219	0.242
Technical and Medical	0.242	0.243	0.257	0.255	0.261	0.253	0.261	0.240
Vocational – technical	0.169	0.184	0.173	0.179	0.184	0.181	0.189	0.162
Vocational – manual	0.048	0.046	0.047	0.048	0.042	0.039	0.038	0.041
High School	0.143	0.149	0.138	0.142	0.148	0.157	0.167	0.192
Incomplete high school	0.174	0.148	0.152	0.140	0.140	0.125	0.126	0.123
Managerial & Professional	0.200	0.210	0.215	0.203	0.215	0.224	0.207	0244
Non-Manual	0.083	0.086	0.087	0.085	0.093	0.093	0.099	0.107
Manual	0.169	0.193	0.174	0.186	0.164	0.179	0.152	0.018
Unskilled	0.060	0.063	0.062	0.063	0.070	0.071	0.067	0.070
Not in labour force	0.258	0.268	0.219	0.263	0.193	0.241	0.201	0.236
Retirement age	0.230	0.179	0.244	0.200	0.266	0.192	0.274	0.164

Appendix 1: Continued.	Round 10 Long.	Round 10 Rep.	Round 13 Long.	Round 13 Rep.	Round 16 Long.	Round 16 Rep.	Round 19 Long.	Round 19 Rep.
Household real income*	7,439	7,699	9,356	10,492	12,405	13,195	14,747	16,485
Equivalised income (OECD)	3,050	3,115	3,853	4,218	5,102	5,161	6,164	6,607
Poverty (=1 if below poverty line)	0.201	0.210	0.088	0.090	0.042	0.046	0.033	0.033
Unemployed	0.045	0.048	0.037	0.044	0.025	0.034	0.030	0.040
Family Size	3.24	3.31	3.16	3.30	3.22	3.41	3.14	3.33
Number of Children 0-7 years	0.21	0.22	0.19	0.21	0.18	0.21	0.19	0.28
Number of Children 7-18 years	0.50	0.53	0.44	0.47	0.39	0.41	0.36	0.37
Chronic disease	0.516	0.497	0.495	0.467	0.476	0.442	0.519	0.453
High blood pressure	0.377	0.353	0.429	0.375	0.438	0.352	0.483	0.359
Health problem (in last 30 days)	0.451	0.427	0.442	0.414	0.434	0.389	0.409	0.341
Smokes	0.331	0.363	0.313	0.359	0.301	0.362	0.276	0.335
Health Evaluation: very good	0.016	0.017	0.012	0.016	0.020	0.024	0.012	0.019
Health Evaluation: good	0.236	0.258	0.236	0.285	0.229	0.299	0.217	0.324
Health Evaluation: average	0.566	0.565	0.565	0.541	0.557	0.528	0.584	0.528
Health Evaluation: poor	0.153	0.135	0.157	0.133	0.165	0.125	0.158	0.114
Health Evaluation: very poor	0.029	0.023	0.030	0.025	0.030	0.024	0.029	0.016
Life Satisfaction: fully	0.047	0.048	0.036	0.045	0.044	0.063	0.056	0.087
Life Satisfaction: rather	0.169	0.172	0.270	0.301	0.321	0.337	0.350	0.370
Life Satisfaction: indifferent	0.236	0.239	0.253	0.244	0.253	0.246	0.258	0.234
Life Satisfaction: not much	0.368	0.367	0.316	0.296	0.263	0.247	0.238	0.226
Life Satisfaction: not at all	0.180	0.175	0.124	0.113	0.119	0.111	0.099	0.084
Attitude: good	0.803	0.798	0.826	0.814	0.831	0.831	n/a	n/a
Understanding: good	0.880	0.889	0.900	0.904	0.907	0.919	n/a	n/a
Observations	*7,309*	*7,309*	*5,316*	*7,187*	*4,350*	*8,521*	*3,715*	*13,610*

Notes: The representative means are weighted using the sampling weights; * deflated to 1992 roubles.

Appendix 2: Socioeconomic characteristics by type of attrition

	Always in	Permanent Attritors	Temporary Attritors – in Round 19	Temporary Attritors – out Round 19
Male	37.3	44.3	43.2	44.6
Female	62.7	55.7	56.8	55.2
Age 18-29	18.8	22.5	27.1	29.0
Age 30-39	18.2	13.8	19.0	15.0
Age 40-49	24.0	17.0	23.7	16.7
Age 50-59	15.6	10.2	14.8	15.2
Age >=60	23.4	36.5	15.5	24.0
Urban	57.2	73.2	72.8	78.5
Rural	34.5	22.3	22.5	15.9
PGT	8.3	4.6	4.7	5.6
Moscow & St. Petersburg	3.6	15.6	19.5	27.9
Northern & North Western	5.1	6.6	7.9	5.4
Central & Central Black-Earth	18.8	18.7	16.1	15.6
Volga-Vyatski & Volga Basin	22.5	14.9	10.3	7.9
North Caucasian	16.8	10.8	16.4	15.0
Ural	15.6	12.6	14.4	10.1
Western Siberian	9.0	10.6	5.9	5.4
Eastern Siberian & Far Eastern	8.8	10.3	9.5	13.7
Married	70.2	58.0	64.8	52.3
Single	12.2	15.4	16.7	22.3
Divorced	6.8	7.8	10.0	22.3
Widowed	10.8	18.9	8.5	14.0
University	22.5	20.5	26.5	27.3
Technical and Medical	26.8	21.0	27.6	22.7
Vocational – technical	18.9	13.5	22.8	18.2
Vocational – manual	4.3	5.9	2.6	4.8
High School	13.1	15.7	13.9	13.4
Incomplete high school	14.4	23.6	6.6	13.6
Managerial & Professional	21.9	17.4	22.4	20.7
Non-Manual	9.7	7.2	7.9	8.6
Manual	18.5	14.6	19.8	17.2
Unskilled	6.4	5.1	7.8	6.5
Not in labour force	25.5	24.7	29.9	28.1
Retirement age	18.0	31.0	12.2	18.8
Poverty (=1 if below poverty line)	20.6	18.8	23.1	21.1
Unemployed	4.0	4.5	5.7	5.7
Equivalised income quintile 1 (lowest 20%)	39.6	37.7	39.4	34.5
Equivalised income quintile 2	26.6	24.3	19.6	23.2
Equivalised income quintile 3	17.3	16.6	18.8	18.9
Equivalised income quintile 4	9.2	11.9	13.3	13.4
Equivalised income quintile 5 (highest 20%)	7.4	9.5	9.1	10.1

Family Size	3.42	3.03	3.43	3.12
Chronic disease	49.7	53.8	49.7	53.0
High blood pressure	37.8	39.4	31.0	37.2
Health problem	42.9	48.9	38.8	44.2
Smokes	28.5	35.6	37.5	37.6
Health Status : very good	1.9	1.1	2.6	1.5
Health Status : good	23.7	22.7	26.9	24.0
Health Status: average	60.0	52.2	60.9	58.4
Health Status: poor	13.1	19.2	8.8	14.2
Health Status: very poor	1.4	4.8	0.8	1.9
Life Satisfaction: fully	3.6	5.4	5.2	5.8
Life Satisfaction: rather	17.5	16.6	16.2	15.6
Life Satisfaction: indifferent	24.1	22.9	24.7	24.0
Life Satisfaction: not much	38.0	35.3	37.6	38.9
Life Satisfaction: not at all	16.8	19.9	16.2	15.8
Observations	*2,946*	*3,128*	*769*	*466*

Appendix 3(a): Female participation equations

	(1) Part11	(2) Part12	(3) Part13	(4) Part14	(5) Part15	(6) Part16	(7) Part17	(8) Part18	(9) Part19
	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE
Age 18 – 29	-0.106	-0.101*	-0.152**	-0.160**	-0.138**	-0.123*	-0.116*	-0.144**	-0.123*
	(0.065)	(0.061)	(0.069)	(0.071)	(0.068)	(0.069)	(0.068)	(0.070)	(0.068)
Age 30 – 39	-0.054	-0.028	-0.048	-0.048	-0.039	-0.034	-0.046	-0.071	-0.046
	(0.058)	(0.053)	(0.062)	(0.066)	(0.065)	(0.066)	(0.067)	(0.069)	(0.068)
Age 40 – 49	-0.032	-0.024	-0.01	-0.042	-0.043	-0.009	0.008	-0.021	-0.004
	(0.051)	(0.051)	(0.058)	(0.064)	(0.063)	(0.063)	(0.064)	(0.067)	(0.066)
Age 50 – 59	-0.013	0.021	0.007	0.014	0.055	0.055	0.058	0.054	0.076
	(0.048)	(0.045)	(0.056)	(0.060)	(0.058)	(0.061)	(0.063)	(0.065)	(0.064)
North/North-west	0.054***	0.088***	0.090***	0.147***	0.139***	0.116***	0.147***	0.159***	0.135***
	(0.015)	(0.020)	(0.028)	(0.028)	(0.034)	(0.040)	(0.042)	(0.042)	(0.044)
Central Black Earth	0.061***	0.107***	0.129***	0.153***	0.179***	0.175***	0.182***	0.173***	0.174***
	(0.013)	(0.017)	(0.021)	(0.025)	(0.027)	(0.030)	(0.032)	(0.033)	(0.035)
Volga Basin	0.089***	0.136***	0.182***	0.227***	0.244***	0.248***	0.247***	0.235***	0.234***
	(0.011)	(0.015)	(0.018)	(0.020)	(0.023)	(0.026)	(0.030)	(0.031)	(0.033)
North Caucasus	0.080***	0.105***	0.139***	0.186***	0.186***	0.211***	0.229***	0.248***	0.194***
	(0.013)	(0.018)	(0.023)	(0.025)	(0.028)	(0.030)	(0.033)	(0.033)	(0.038)
Urals	0.079***	0.138***	0.165***	0.208***	0.232***	0.238***	0.245***	0.229***	0.205***
	(0.012)	(0.014)	(0.018)	(0.021)	(0.023)	(0.026)	(0.030)	(0.032)	(0.035)
Western Siberia	0.025	0.050**	0.074***	0.113***	0.113***	0.117***	0.123***	0.096**	0.091**
	(0.019)	(0.023)	(0.027)	(0.029)	(0.033)	(0.036)	(0.038)	(0.040)	(0.041)
E.Siberia/Far East	0.021	0.056**	0.074***	0.122***	0.150***	0.127***	0.156***	0.092**	0.101**
	(0.019)	(0.022)	(0.027)	(0.028)	(0.030)	(0.035)	(0.036)	(0.040)	(0.041)
Rural	0.067***	0.103***	0.105***	0.125***	0.139***	0.130***	0.137***	0.157***	0.146***
	(0.012)	(0.014)	(0.017)	(0.019)	(0.020)	(0.021)	(0.022)	(0.022)	(0.023)
Urban Type (PGT)	0.031*	0.042*	0.038	0.026	0.071**	0.084**	0.120***	0.115***	0.113***
	(0.019)	(0.025)	(0.032)	(0.036)	(0.035)	(0.036)	(0.036)	(0.037)	(0.038)
Single	0.01	0.026	-0.023	-0.023	-0.019	-0.031	-0.050*	-0.044	-0.072**
	(0.016)	(0.019)	(0.024)	(0.026)	(0.027)	(0.029)	(0.030)	(0.030)	(0.030)
Divorced	-0.02	-0.019	-0.035	-0.027	-0.009	-0.045	-0.064**	-0.051*	-0.065**
	(0.018)	(0.022)	(0.025)	(0.026)	(0.027)	(0.029)	(0.030)	(0.030)	(0.030)
Widowed	-0.037**	-0.045**	-0.059**	-0.096***	-0.093***	-0.081***	-0.084***	-0.089***	-0.115***
	(0.018)	(0.021)	(0.023)	(0.025)	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)
University	0.053***	0.104***	0.129***	0.103***	0.124***	0.099***	0.117***	0.144***	0.152***
	(0.015)	(0.017)	(0.021)	(0.025)	(0.026)	(0.029)	(0.030)	(0.030)	(0.031)
Tech/Medical	0.042***	0.092***	0.090***	0.071***	0.079***	0.069**	0.074***	0.097***	0.114***
	(0.014)	(0.017)	(0.021)	(0.024)	(0.025)	(0.027)	(0.028)	(0.028)	(0.028)
Tech/Vocational	0.014	0.057***	0.054**	0.065**	0.094***	0.068**	0.069**	0.087**	0.117***
	(0.019)	(0.019)	(0.025)	(0.028)	(0.029)	(0.032)	(0.034)	(0.034)	(0.034)
Vocational	0.022	0.089***	0.075**	0.042	0.051	-0.029	-0.022	-0.002	0.036
	(0.025)	(0.022)	(0.033)	(0.042)	(0.045)	(0.051)	(0.052)	(0.052)	(0.051)
High School	-0.028	0.005	-0.017	-0.005	-0.005	-0.046	-0.037	-0.03	0.023
	(0.021)	(0.022)	(0.027)	(0.029)	(0.030)	(0.032)	(0.033)	(0.033)	(0.033)
Managerial/Prof	-0.086**	-0.082**	-0.12***	-0.037	-0.081*	-0.069	-0.080*	-0.088*	-0.099**
	(0.038)	(0.041)	(0.045)	(0.041)	(0.045)	(0.044)	(0.046)	(0.046)	(0.045)
Non-manual	-0.072*	-0.068	-0.090**	-0.051	-0.06	-0.042	-0.068	-0.058	-0.085*
	(0.040)	(0.042)	(0.045)	(0.043)	(0.046)	(0.046)	(0.047)	(0.048)	(0.046)
Manual	-0.074	-0.085*	-0.15***	-0.082	-0.078	-0.079	-0.072	-0.06	-0.048
	(0.047)	(0.051)	(0.055)	(0.051)	(0.054)	(0.054)	(0.054)	(0.054)	(0.054)
Not in labour force	-0.10***	-0.11***	-0.12***	-0.10**	-0.14***	-0.12***	-0.14***	-0.14***	-0.14***

	(0.037)	(0.041)	(0.043)	(0.041)	(0.045)	(0.044)	(0.045)	(0.045)	(0.044)
Retirement age	-0.112*	-0.104*	-0.146**	-0.105	-0.142**	-0.132*	-0.161**	-0.184**	-0.20***
	(0.064)	(0.063)	(0.071)	(0.071)	(0.071)	(0.071)	(0.072)	(0.074)	(0.071)
Poverty (=1 if below poverty line)	-0.07***	-0.061**	-0.027	-0.019	-0.049	-0.034	-0.043	-0.023	-0.025
	(0.025)	(0.026)	(0.027)	(0.029)	(0.030)	(0.031)	(0.031)	(0.031)	(0.031)
Income Q2	-0.022	-0.011	-0.004	0.007	-0.015	0.009	0.002	0.012	0.007
	(0.019)	(0.021)	(0.023)	(0.025)	(0.026)	(0.027)	(0.027)	(0.028)	(0.028)
Income Q3	-0.062**	-0.042	-0.021	-0.011	-0.063**	-0.023	-0.037	-0.024	-0.012
	(0.025)	(0.026)	(0.027)	(0.029)	(0.031)	(0.031)	(0.031)	(0.031)	(0.032)
Income Q4	-0.044	-0.073**	-0.078**	-0.068*	-0.10***	-0.066*	-0.092**	-0.077**	-0.076**
	(0.028)	(0.032)	(0.034)	(0.035)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)
Income Q5	-0.10***	-0.13***	-0.10***	-0.11***	-0.11**	-0.10**	-0.13***	-0.12***	-0.13***
	(0.035)	(0.038)	(0.038)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Family size	0.002	0.008	0.007	0.005	0.009	0.006	0.007	0.003	-0.001
	(0.005)	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Have children	0.019	0.016	0.02	0.019	0.027	0.059**	0.059**	0.074***	0.064**
	(0.015)	(0.019)	(0.022)	(0.024)	(0.024)	(0.025)	(0.026)	(0.026)	(0.026)
Health very poor	-0.057	-0.195*	-0.194**	-0.252**	-0.248**	-0.30***	-0.24***	-0.26***	-0.29***
	(0.067)	(0.111)	(0.098)	(0.101)	(0.100)	(0.093)	(0.090)	(0.089)	(0.081)
Health poor	-0.002	-0.075	-0.055	-0.095	-0.08	-0.087	-0.087	-0.1	-0.155*
	(0.045)	(0.079)	(0.074)	(0.083)	(0.086)	(0.086)	(0.084)	(0.086)	(0.083)
Health average	0.027	-0.043	-0.014	-0.058	-0.027	-0.054	-0.049	-0.052	-0.092
	(0.045)	(0.062)	(0.064)	(0.070)	(0.076)	(0.077)	(0.078)	(0.080)	(0.079)
Health good	0.011	-0.1	-0.049	-0.104	-0.071	-0.081	-0.048	-0.053	-0.11
	(0.041)	(0.080)	(0.070)	(0.080)	(0.082)	(0.082)	(0.080)	(0.083)	(0.081)
Health problem in last 30 days	0.022*	0.028**	0.029*	0.014	0.032*	0.036*	0.018	0.018	0.023
	(0.011)	(0.014)	(0.016)	(0.017)	(0.018)	(0.019)	(0.020)	(0.020)	(0.019)
High blood pres.	0.007	0.030**	0.02	0.037**	0.027	0.009	0.02	0.014	0.004
	(0.011)	(0.014)	(0.016)	(0.017)	(0.018)	(0.019)	(0.020)	(0.020)	(0.020)
Not at all satisfied	0.014	0.041	0.057*	0.065*	0.041	0.027	0.045	0.024	0.024
	(0.024)	(0.029)	(0.033)	(0.037)	(0.041)	(0.044)	(0.046)	(0.047)	(0.048)
Less than satisfied	0.009	0.031	0.070**	0.091**	0.063	0.064	0.090**	0.072*	0.057
	(0.023)	(0.029)	(0.032)	(0.036)	(0.039)	(0.041)	(0.043)	(0.044)	(0.044)
Indifferent	0.015	0.027	0.062*	0.082**	0.051	0.043	0.063	0.066	0.06
	(0.023)	(0.029)	(0.031)	(0.034)	(0.039)	(0.042)	(0.043)	(0.044)	(0.045)
Rather satisfied	0.028	0.036	0.073**	0.110***	0.086**	0.063	0.078*	0.067	0.052
	(0.022)	(0.028)	(0.031)	(0.033)	(0.038)	(0.042)	(0.044)	(0.045)	(0.046)
Attitude	0.049***	0.059***	0.090***	0.090***	0.109***	0.096***	0.082***	0.090***	0.094***
	(0.017)	(0.021)	(0.023)	(0.025)	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)
Understanding	0.009	0.058**	0.064***	0.095***	0.064**	0.055*	0.058**	0.046	0.022
	(0.018)	(0.023)	(0.025)	(0.027)	(0.028)	(0.028)	(0.029)	(0.029)	(0.029)
Sample Size	*3796*	*3796*	*3796*	*3796*	*3796*	*3796*	*3796*	*3796*	*3796*
Log Likelihood	-1248.86	-1581.07	-1852.44	-2023	-2133.89	-2258.62	-2336.24	-2333.89	-2366.28

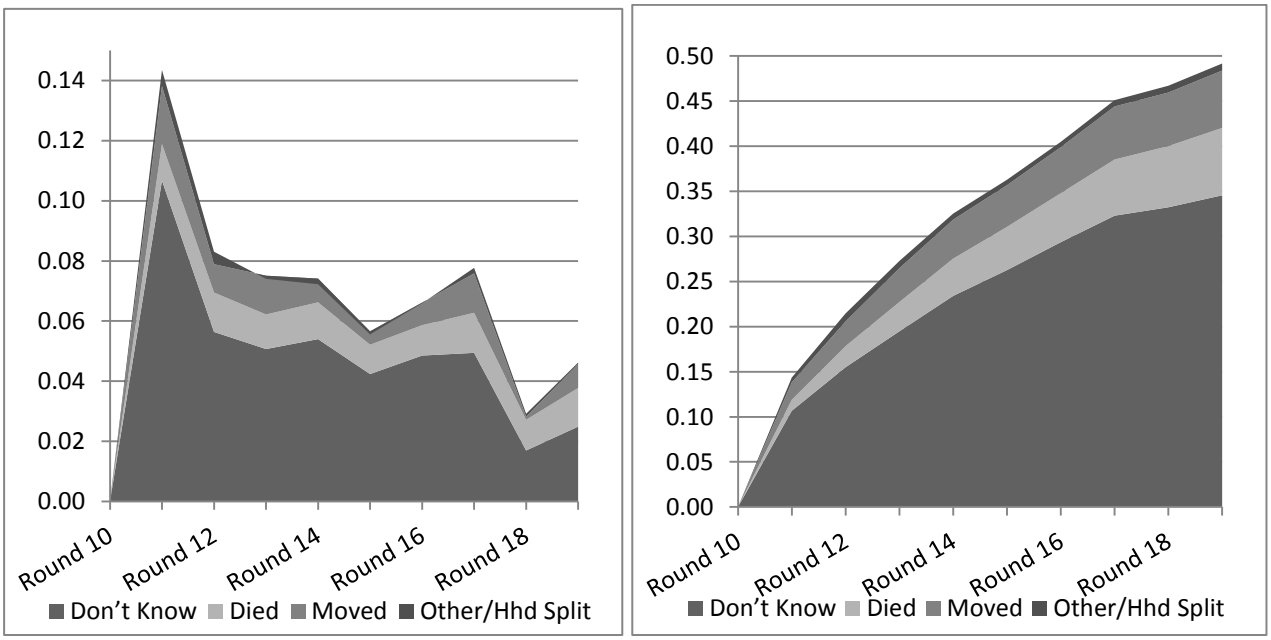
Notes (also for Appendix 3b): Robust standard errors, clustered by family, in parentheses;
* / ** /*** denotes respectively 10% / 5% / 1% level of significance; coefficients show the marginal effects evaluated at the mean values of other variables.
The excluded categories for the dummy variables are: Age 60+, Moscow & St. Petersburg, urban, married, incomplete high school, unskilled occupation, income Q1, very good health, very satisfied with life.

Appendix 3(b): Male participation equations

	(1) Part11	(2) Part12	(3) Part13	(4) Part14	(5) Part15	(6) Part16	(7) Part17	(8) Part18	(9) Part19
	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE	Coeff/SE
Age 18 – 29	-0.18***	-0.18***	-0.23***	-0.17***	-0.16***	-0.16***	-0.14***	-0.15***	-0.11**
	(0.045)	(0.047)	(0.049)	(0.049)	(0.049)	(0.047)	(0.046)	(0.046)	(0.046)
Age 30 – 39	-0.088**	-0.071*	-0.14***	-0.14***	-0.16***	-0.13***	-0.111**	-0.097**	-0.071
	(0.040)	(0.041)	(0.046)	(0.047)	(0.046)	(0.045)	(0.045)	(0.044)	(0.044)
Age 40 – 49	-0.068*	-0.045	-0.12***	.125***	-0.104**	-0.073*	-0.046	-0.06	-0.034
	(0.036)	(0.038)	(0.044)	(0.044)	(0.043)	(0.043)	(0.042)	(0.042)	(0.042)
Age 50 – 59	-0.076*	-0.071*	-0.094**	-0.108**	-0.081*	-0.081*	-0.066	-0.077*	-0.044
	(0.039)	(0.041)	(0.045)	(0.045)	(0.045)	(0.044)	(0.043)	(0.043)	(0.043)
North/North-west	0.074***	0.059*	0.069	0.129***	0.181***	0.100*	0.123**	0.118**	0.06
	(0.022)	(0.036)	(0.043)	(0.044)	(0.045)	(0.053)	(0.056)	(0.057)	(0.060)
Central Black Earth	0.083***	0.101***	0.135***	0.169***	0.212***	0.191***	0.178***	0.172***	0.158***
	(0.019)	(0.027)	(0.032)	(0.034)	(0.036)	(0.040)	(0.043)	(0.045)	(0.046)
Volga Basin	0.126***	0.149***	0.194***	0.218***	0.272***	0.258***	0.248***	0.246***	0.229***
	(0.016)	(0.024)	(0.029)	(0.032)	(0.033)	(0.038)	(0.042)	(0.043)	(0.044)
North Caucasus	0.103***	0.082***	0.135***	0.181***	0.215***	0.232***	0.219***	0.252***	0.202***
	(0.019)	(0.032)	(0.035)	(0.036)	(0.039)	(0.041)	(0.045)	(0.045)	(0.048)
Urals	0.088***	0.154***	0.175***	0.231***	0.265***	0.241***	0.231***	0.234***	0.215***
	(0.018)	(0.022)	(0.029)	(0.030)	(0.033)	(0.038)	(0.042)	(0.044)	(0.046)
Western Siberia	0.063***	0.019	0.051	0.087**	0.119***	0.132***	0.142***	0.145***	0.123**
	(0.022)	(0.037)	(0.041)	(0.042)	(0.045)	(0.047)	(0.049)	(0.051)	(0.052)
E.Siberia/Far East	0.04	0.037	0.048	0.143***	0.161***	0.176***	0.150***	0.117**	0.103**
	(0.025)	(0.035)	(0.041)	(0.037)	(0.041)	(0.043)	(0.048)	(0.050)	(0.051)
Rural	0.083***	0.134***	0.153***	0.196***	0.200***	0.174***	0.179***	0.198***	0.198***
	(0.015)	(0.018)	(0.021)	(0.022)	(0.023)	(0.025)	(0.025)	(0.026)	(0.026)
Urban type	0.027	0.108***	0.075**	0.118***	0.083**	0.062	0.082*	0.080*	0.105**
	(0.028)	(0.030)	(0.036)	(0.038)	(0.042)	(0.045)	(0.045)	(0.045)	(0.045)
Single	0.018	0.004	-0.005	-0.027	-0.017	-0.012	-0.024	-0.009	-0.033
	(0.022)	(0.028)	(0.032)	(0.035)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Divorced	-0.080**	-0.024	-0.029	-0.072	-0.028	-0.073	-0.08	-0.078	-0.108**
	(0.039)	(0.041)	(0.045)	(0.049)	(0.049)	(0.050)	(0.049)	(0.049)	(0.048)
Widowed	-0.068	-0.059	-0.14***	-0.20***	-0.23***	-0.20***	-0.21***	-0.20***	-0.23***
	(0.044)	(0.047)	(0.052)	(0.055)	(0.055)	(0.055)	(0.054)	(0.055)	(0.054)
University	0.079***	0.144***	0.171***	0.223***	0.206***	0.191***	0.191***	0.212***	0.180***
	(0.021)	(0.024)	(0.029)	(0.031)	(0.034)	(0.037)	(0.039)	(0.040)	(0.042)
Tech/Medical	0.047**	0.119***	0.152***	0.215***	0.212***	0.180***	0.182***	0.213***	0.182***
	(0.023)	(0.024)	(0.028)	(0.029)	(0.032)	(0.035)	(0.037)	(0.038)	(0.040)
Tech/Vocational	0.054**	0.131***	0.136***	0.197***	0.190***	0.172***	0.193***	0.225***	0.199***
	(0.022)	(0.024)	(0.029)	(0.031)	(0.034)	(0.036)	(0.037)	(0.037)	(0.039)
Vocational	0.039	0.104***	0.087**	0.129***	0.127***	0.087**	0.076*	0.113**	0.059
	(0.026)	(0.027)	(0.035)	(0.037)	(0.040)	(0.043)	(0.045)	(0.045)	(0.047)
High School	0.045**	0.086***	0.090***	0.115***	0.126***	0.059	0.101**	0.117***	0.074*
	(0.023)	(0.026)	(0.031)	(0.034)	(0.036)	(0.040)	(0.040)	(0.041)	(0.042)
Managerial/Prof	-0.05	-0.062	-0.026	-0.056	-0.017	-0.062	-0.038	-0.007	0.006
	(0.040)	(0.046)	(0.046)	(0.050)	(0.050)	(0.051)	(0.051)	(0.050)	(0.050)
Non-manual	-0.043	-0.062	-0.052	-0.105	-0.057	-0.06	-0.039	-0.071	0.03
	(0.055)	(0.067)	(0.069)	(0.073)	(0.071)	(0.070)	(0.072)	(0.069)	(0.070)
Manual	-0.028	0.007	0.058	0.041	0.089**	0.06	0.044	0.059	0.059
	(0.031)	(0.037)	(0.037)	(0.041)	(0.041)	(0.042)	(0.043)	(0.042)	(0.043)
Not in labour force	-0.05	-0.04	-0.035	-0.048	0.004	0.003	-0.015	-0.001	0.02

	(0.034)	(0.039)	(0.041)	(0.043)	(0.044)	(0.044)	(0.045)	(0.044)	(0.044)
Retirement age	-0.102*	-0.062	-0.123**	-0.147**	-0.073	-0.131**	-0.130**	-0.131**	-0.135**
	(0.054)	(0.054)	(0.059)	(0.061)	(0.061)	(0.060)	(0.058)	(0.058)	(0.058)
Poverty (=1 if below poverty line)	-0.038	-0.031	-0.014	-0.092**	-0.11***	-0.08**	-0.12***	-0.12***	-0.094**
	(0.027)	(0.031)	(0.033)	(0.038)	(0.038)	(0.038)	(0.037)	(0.037)	(0.037)
Income Q2	0.01	0.016	-0.001	-0.046	-0.053	-0.013	-0.043	-0.04	-0.045
	(0.023)	(0.028)	(0.030)	(0.034)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Income Q3	-0.023	0.006	0.019	-0.033	-0.044	-0.022	-0.04	-0.039	-0.033
	(0.027)	(0.031)	(0.033)	(0.037)	(0.038)	(0.039)	(0.038)	(0.039)	(0.038)
Income Q4	-0.003	-0.011	-0.012	-0.021	-0.068	-0.059	-0.096**	-0.075*	-0.097**
	(0.029)	(0.036)	(0.039)	(0.042)	(0.043)	(0.043)	(0.042)	(0.044)	(0.043)
Income Q5	-0.088**	-0.100**	-0.091**	-0.114**	-0.072	-0.046	-0.108**	-0.101**	-0.083*
	(0.037)	(0.041)	(0.043)	(0.046)	(0.046)	(0.046)	(0.045)	(0.045)	(0.045)
Family size	0.013*	0.014*	0.021**	0.019*	0.033***	0.028***	0.032***	0.028***	0.024**
	(0.007)	(0.008)	(0.008)	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)
Have children	0.000	-0.003	-0.003	0.007	0.014	0.047	0.01	0.022	0.021
	(0.019)	(0.024)	(0.027)	(0.028)	(0.029)	(0.029)	(0.029)	(0.030)	(0.030)
Health very poor	-0.159	-0.274**	-0.39***	-0.39***	-0.35***	-0.34***	-0.34***	-0.38***	-0.35***
	(0.110)	(0.115)	(0.107)	(0.099)	(0.097)	(0.091)	(0.081)	(0.068)	(0.071)
Health poor	-0.043	-0.073	-0.194**	-0.188**	-0.24***	-0.22***	-0.21***	-0.22***	-0.21***
	(0.061)	(0.068)	(0.082)	(0.083)	(0.078)	(0.077)	(0.072)	(0.070)	(0.068)
Health average	-0.024	0.016	-0.093	-0.08	-0.09	-0.082	-0.096	-0.125*	-0.1
	(0.048)	(0.054)	(0.065)	(0.070)	(0.069)	(0.071)	(0.070)	(0.069)	(0.068)
Health good	-0.065	0.002	-0.101	-0.113	-0.112	-0.105	-0.122*	-0.153**	-0.104
	(0.052)	(0.053)	(0.068)	(0.071)	(0.070)	(0.071)	(0.069)	(0.068)	(0.066)
Health problem in last 30 days	0.005	0.017	-0.002	0.009	0.034	0.011	0.003	0.01	0.002
	(0.016)	(0.019)	(0.021)	(0.023)	(0.023)	(0.024)	(0.024)	(0.024)	(0.025)
High blood pres.	-0.003	-0.007	-0.009	-0.022	-0.041*	-0.038	-0.038	-0.031	-0.03
	(0.016)	(0.019)	(0.021)	(0.022)	(0.023)	(0.023)	(0.023)	(0.024)	(0.024)
Not at all satisfied	-0.012	0.064*	0.071*	0.102**	0.077	0.069	0.066	0.046	0.041
	(0.036)	(0.035)	(0.041)	(0.044)	(0.049)	(0.051)	(0.052)	(0.053)	(0.054)
Less than satisfied	-0.02	0.029	0.043	0.061	0.06	0.035	0.056	0.039	0.032
	(0.032)	(0.036)	(0.041)	(0.044)	(0.047)	(0.048)	(0.048)	(0.048)	(0.049)
Indifferent	-0.024	0.039	0.041	0.072*	0.055	0.043	0.071	0.059	0.049
	(0.034)	(0.035)	(0.041)	(0.043)	(0.047)	(0.048)	(0.048)	(0.049)	(0.050)
Rather satisfied	0.01	0.046	0.046	0.091**	0.088*	0.073	0.100**	0.111**	0.095*
	(0.032)	(0.035)	(0.041)	(0.044)	(0.047)	(0.049)	(0.049)	(0.049)	(0.051)
Attitude	0.018	0.047**	0.050**	0.054**	0.046*	0.049*	0.051**	0.032	0.042*
	(0.017)	(0.022)	(0.024)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Understanding	0.039	0.053*	0.037	0.02	0.055	0.038	0.027	0.01	-0.003
	(0.026)	(0.029)	(0.031)	(0.033)	(0.034)	(0.034)	(0.035)	(0.035)	(0.036)
Sample Size	*2760*	*2760*	*2760*	*2760*	*2760*	*2760*	*2760*	*2760*	*2760*
Log Likelihood	-1091.2	-1361.59	-1526.02	-1602.11	-1642.67	-1692.14	-1726.95	-1716.32	-1718.12

Figure 1: Attrition from the longitudinal sample



a) Round by round attrition hazard **b) Cumulative round by round attrition rate**

Figure 2: Participation rates over time conditional on round 10 characteristics

Figure 2(a): by **Region**

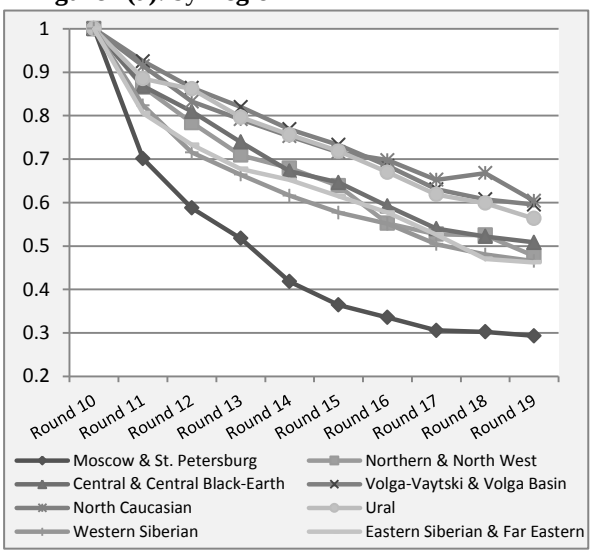


Figure 2(b): by **Settlement Type**

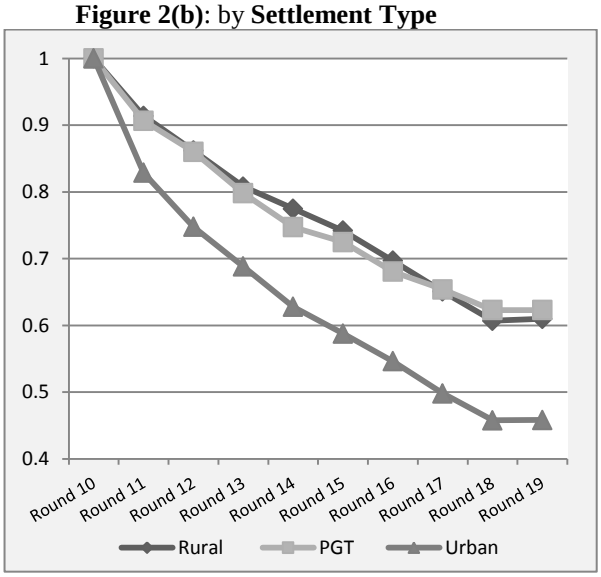


Figure 2(c): by **Age Group**

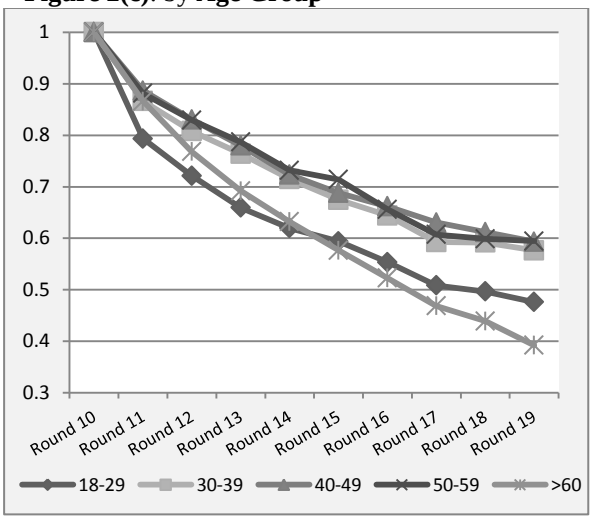


Figure 2(d): by **Gender**

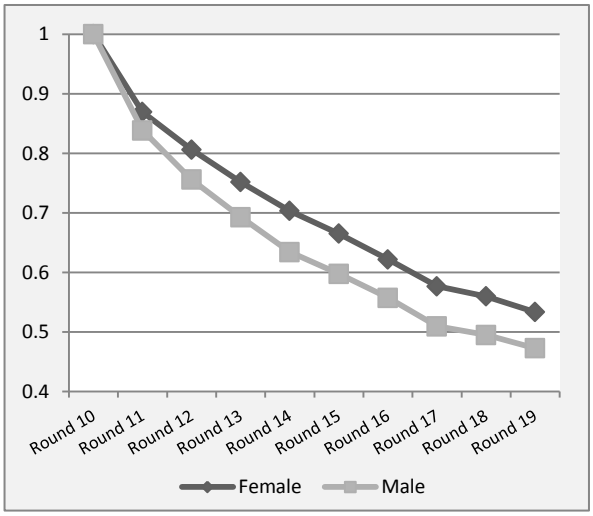


Figure 2(e): by **Marital Status**

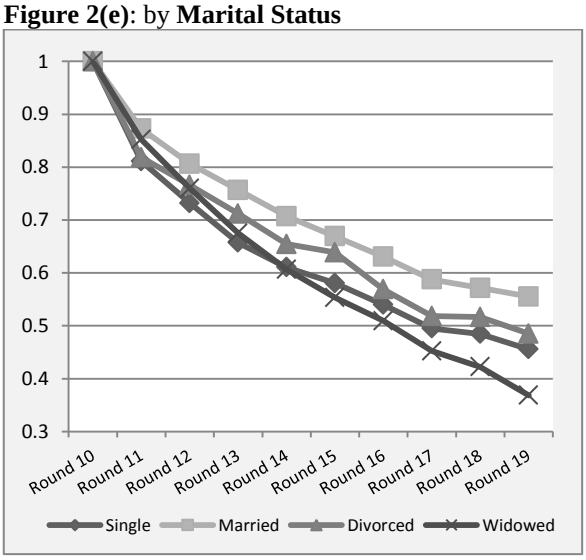


Figure 2(f): by **Education Level**

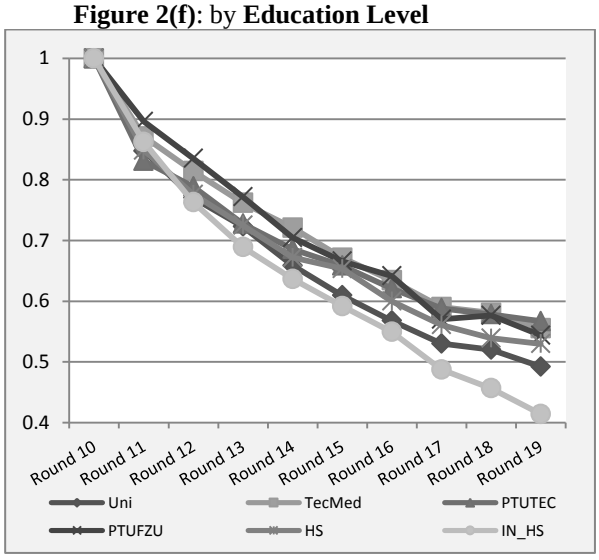


Figure 2(g): by **Health Evaluation**

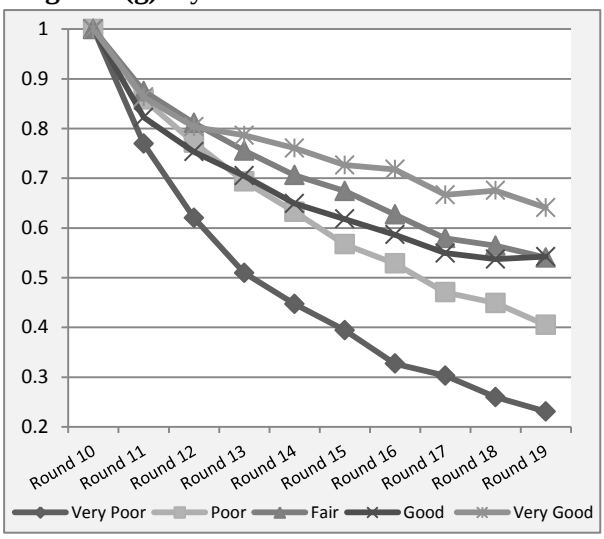


Figure 2(h): by **Health Problem in last 30** days

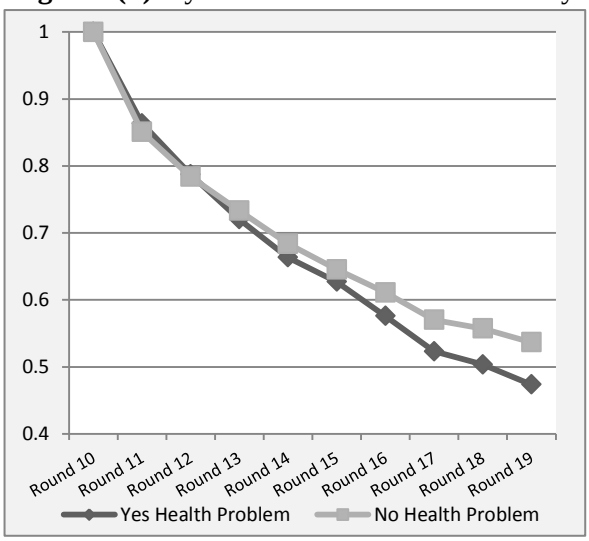


Figure 2(i): by **Ever Told High Blood Pressure**

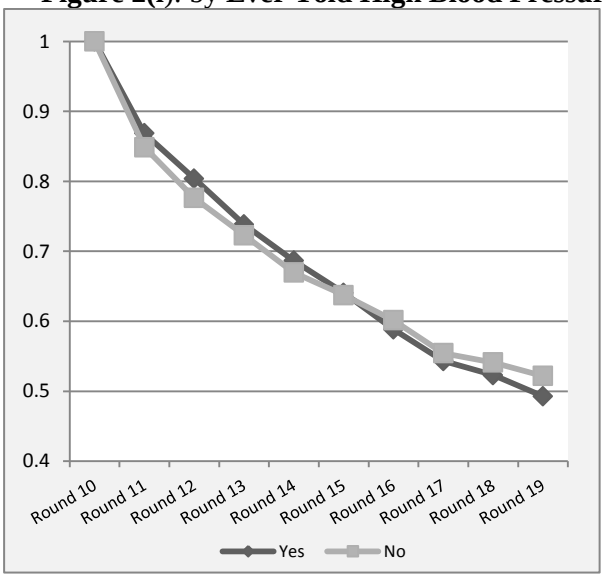


Figure 2(j): by **Life Satisfaction**

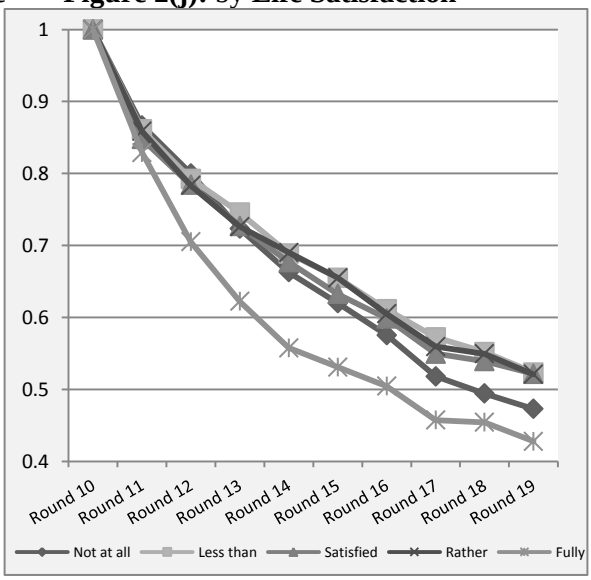


Figure 2(k): by **Hhd Income Quintile**

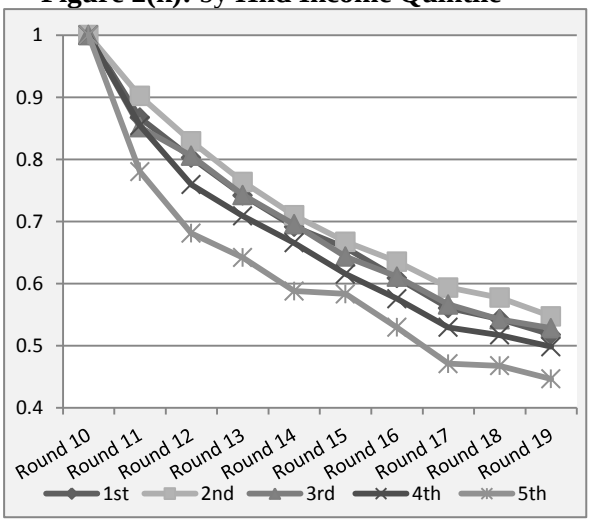


Figure 2(l): by **Unemployment Status**