

# Sample attrition and labor force dynamics: Evidence from the Spanish labor force survey

Sergi Jiménez-Martín<sup>1</sup>, Franco Peracchi<sup>2</sup>

<sup>1</sup> Departamento de Economía, Universidad Carlos III de Madrid, Madrid 126, 28903 Getafe, Spain

<sup>2</sup> Facoltà di Economia, Università di Roma Tor Vergata, 00133 Roma, Italy

**Abstract.** In this paper we examine a number of issues that arise in investigating labor force dynamics using the Spanish Labor Force Survey (EPA). These issues are by no means specific to the Spanish case and apply to most European-style labor force surveys. Our main conclusions may be summarized as follows. First, survey nonresponse cannot be neglected. Second, the EPA tends to underestimate employment and participation of high-educated young people, and to overestimate those of the low-educated elderly. Finally, we find little evidence that attrition causes important selection biases in estimating quarterly transition probabilities.

**JEL Classification:** C23, C81, J21

**Key words:** Nonresponse, attrition, panel data, labor force survey, EPA

## 1 Introduction

Understanding labor market behavior is of fundamental importance in modern economies. To this purpose, a labor force survey is typically carried out in order to study participation, unemployment, search behavior and labor force transitions, especially entry and exit into and from employment. However, phenomena such as attrition, nonresponse and recall errors may seriously affect what we can learn from survey data.<sup>1</sup>

---

We thank Paqui Martínez for excellent research assistance. We also thank participants at the 2nd Encuentro de Economía Aplicada, Zaragoza, 1999, and an anonymous referee for helpful comments on an earlier version. This paper is part of a research project financed by Fundación BBV under the title “Reforming the Spanish Social Security system”.

<sup>1</sup> See van den Berg and Lindeboom (1998) or Zabel (1998) for examples of the consequences of attrition on panel data analysis, and Paull (1997) for the consequences of recall errors on the study of dynamic labor market behavior.

In this paper we examine a number of issues that arise in investigating labor force dynamics using the Spanish Labor Force Survey (*Encuesta de Población Activa*, EPA). The EPA is the main source of information on the employment behavior of the Spanish population. Our analysis is based on micro-data spanning from the second quarter of 1987 to the fourth quarter of 1998. Although we focus on the EPA, some of the issues that we raise go beyond the Spanish case and apply to the kind of labor force surveys which are carried out in most European countries.

Besides being a source of cross-sectional information on labor force participation, the EPA is also a source of longitudinal information on labor force dynamics. There are essentially two ways in which this longitudinal information may be obtained. One is to exploit the rotating nature of the EPA which makes it possible to match individuals across waves of the survey, the other is to exploit the retrospective questions on labor force behavior contained in the survey. We focus on the former because is the one most commonly used.<sup>2</sup>

We begin by describing the problem of sample attrition and re-entry in the EPA. Attrition may have important consequences since the sample may lose representativeness of the target population.<sup>3</sup> We distinguish two types of attrition: monotone and nonmonotone. In the first case, an individual who leaves the sample is never observed again. In the second case, an individual leaves the sample only temporarily and is observed again some period later. A related phenomenon is entry of new individuals into the sample after the first interview. New entry and re-entry (non-monotone attrition) may help compensating the consequences of nonresponse in the EPA. We estimate a simple model of attrition and re-entry that captures the evolution of response patterns in the EPA.

In the second part of the paper we look at the problems that attrition poses for the estimation of labor force dynamics. We first compare annual and quarterly transitions without taking attrition into account. We then evaluate the degree of uncertainty caused by sample attrition by constructing bounds within which the population transition probabilities must lie. We show that this kind of uncertainty is much more important than the one caused by sampling variability. Finally, we perform nonparametric tests of attrition by exploiting the differences between transition probabilities for full-time respondents (those who stay in for the full cycle of interviews) and the other sample participants. We also exploit the patterns of response between successive interviews to perform a regression-based test of attrition. Taking the evidence as a whole, we find no evidence that attrition causes important selection biases in transition probabilities.

The plan of the paper is as follows. Section 2 describes the data. Section 3 describes the dynamics of the EPA sample. Section 4 examines the issue of attrition. Section 5 studies the effect of attrition on individual labor force transition. Finally, Sect. 6 contains some conclusions.

---

<sup>2</sup> Examples include Bover et al. (1996) and Alba-Ramírez (1997, 1998).

<sup>3</sup> Villagarcía (1995) provides an excellent illustration of the consequences of nonresponse on labor force figures.

## 2 The data

The EPA is a rotating quarterly survey carried out by the Spanish National Statistical Institute (*Instituto Nacional de Estadística*, INE).<sup>4</sup> The target population consists of all individuals living in private households. Collective households (hospitals, group-quarters, etc.) and secondary residences are excluded from the sample. Interviews are carried out throughout the quarter and not during a specific week. The primary sampling units are the census tracts, while the households (in fact, the street addresses) are the secondary sampling units. The EPA currently considers 3204 census tracts and, on average, 20 households within each of them. Thus, the planned sample size currently consists of approximately 64,000 households with approximately 150,000 adult (aged 16+) individuals.

The survey's rotation scheme is such that every new rotation group stays in the survey for six quarters (one and a half year). This implies that in any two consecutive waves there are 5 overlapping rotation groups. Because of this rotation scheme, in the absence of attrition and new entry, 5/6 of the sample in any two consecutive quarters could be matched and, in any quarter, those in rotation groups 1–2 could be matched with those in rotation groups 5–6 the same quarter of the following year. The procedure that we used to match individuals across waves of the survey is described in the Appendix.

The INE distinguishes between eligible and ineligible persons. Eligible persons are all adult household members who are present or temporarily absent for less than a quarter. Household members who are away for more than a quarter are no longer considered eligible, but they re-enter the sample upon return the household. Those who join a household after the first interview are added to the list of eligible persons, whereas those who leave the household are not followed.

The questionnaire is submitted to a single household respondent who answers for all the persons living in the household. The household respondent may change between successive interviews. The fact that interviews in the EPA are not strictly personal may help explain the moderate attrition rates. It may also help reduce the importance of endogenous attrition for certain groups of the population (for instance, prime-age individuals or individuals near retirement) but, unfortunately, it may introduce measurement errors, especially in retrospective questions. The nature and extent of these measurement errors may depend on the household size and the relationship between the household respondent and each other person.

INE follows standard procedures in order to reduce nonresponse rates due to non-contact, physical incapacity or inability to respond. In the first wave, absent or inaccessible households are re-visited during a period of six weeks until the information is obtained. Households that remain inaccessible are visited again the following quarter. Households that refuse to respond to their first interview are randomly replaced by other dwellings in the same census tract. Refusal to respond to subsequent interviews is treated somewhat differently. In this case, nonresponding households are not generally replaced and the data obtained in the former interview are simply replicated. However, the longitudinal identifiers are changed to

<sup>4</sup> See Jiménez-Martín and Peracchi (1999) for a description of the characteristics of the EPA, and the tracing, checking and data evaluation methodologies adopted by the INE.

avoid unrealistic measures of variation.<sup>5</sup> These households are revisited during the following waves to obtain responses. In case of further nonresponse, data are no longer imputed. For item nonresponse or inconsistencies, an automatic imputation procedure is used (*Detección y Imputación Automática, DIA*).<sup>6</sup>

The survey has been conducted since 1964, and publicly released cross-sectional files are available from 1976. The 1976 questionnaire was modified in 1987 when a set of retrospective questions was introduced, and again in the first quarter of 1992. In both cases, the lengthening of the questionnaire led to increased nonresponse rates.<sup>7</sup> Major changes to the survey have been introduced in the first quarter of 1999. We decided to end our analysis with the last quarter of 1998 because of the insufficient data to fully assess the consequences of the latter changes.

### 3 Sample dynamics

Figure 1 analyzes the distribution of household response rates during the period from 1987:II to 1997:IV, distinguishing between the first interview and the others (interviews 2–6).<sup>8</sup> The first column compares the planned sample, the potential sample (number of households found) and the achieved sample size (number of households interviewed). There are three noticeable features of the data. First, whereas the potential sample size increased by 10 % over the period, the achieved sample size increased much less. Second, the data contain two important discontinuities, one in the third quarter of 1988 and the other in the first quarter of 1992, due to a sharp decrease respectively in the number of households interviewed and in the number of potential households. Third, the achieved sample size displays strong seasonal fluctuations and tends to be higher in the second quarter (April–June) and lower in the third quarter (July–September) of each year.

The central column of the figure shows the time pattern of the fraction of ineligible households and of negatives (refusals to respond). Notice the importance of ineligible households in the first interview (central panel) and its peculiar saw-tooth profile. The sharp drops in 1987, 1992 and 1997 follow the updates of the sampling frame, whereas the monotonic increase of the fraction of ineligible households between successive updates reflects the progressive aging of frame. After the first interview, the fraction of ineligible households is much lower and displays no trend (bottom panel), although it is characterized by seasonal variation.

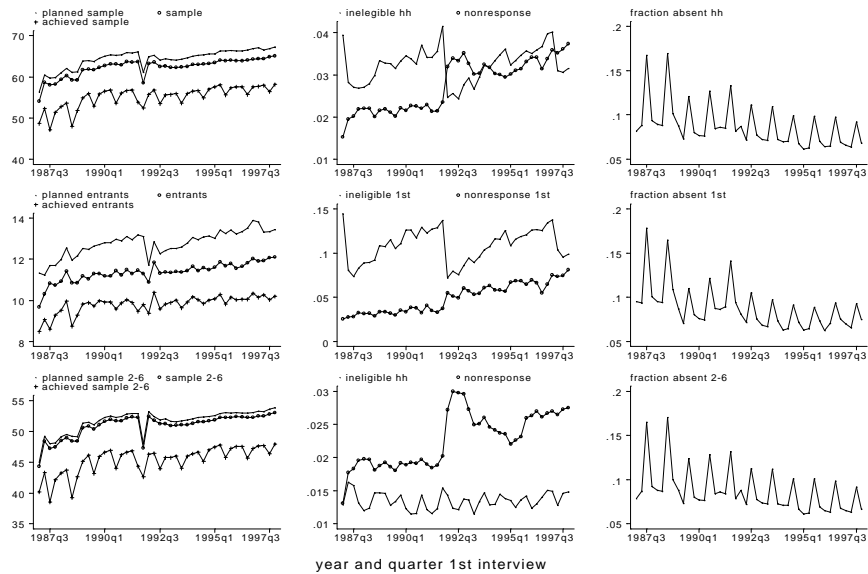
The fraction of negatives shows a clear upward trend. It starts at 2% in 1987 and reaches 3.7% in 1997. As expected, refusal to respond is more important in the first interview (central panel) than in the others (bottom panel). Notice the temporary increase for interviews 2–6 after the change of questionnaire in the first quarter of 1992. On the other hand, the effect of the change on those who were successfully interviewed in wave 1 has been modest. As we shall see below, the

<sup>5</sup> We exclude these data from the analyses in Sects. 4 and 5 below.

<sup>6</sup> See INE (1994) for a comprehensive description of the DIA imputation methodology.

<sup>7</sup> See INE (1992) for a detailed explanation.

<sup>8</sup> The data of this section has been obtained from the series *Incidencia de los Trabajos de Campo en la Encuesta de Poblacion Activa, 19xx*. INE(1999) analyzes the data for 1997, the last year available.



**Fig. 1.** Planned and achieved sample size, and causes of nonresponse: 1987–1998 (Data is taken from INE, *Incidencia de los Trabajos de Campo en la Encuesta de Poblacion Activa, 19xx.*)

steady increase in the fraction of refusals to respond can be partly explained by changes in the composition of the population by education, as the more educated people tend to refuse to participate to the survey more often than the less educated ones. The different value of time for the two groups may help explain the differences in response patterns.

Finally, the last column of Fig. 1 shows the behavior of the fraction of absent household. This fraction displays a strong seasonal pattern around a downward trend. The trend reflects INE’s efforts to increase the number of attempts to contact absent households, while the seasonal effect in the third quarter is due to the summer vacations.

To summarize, INE tried to counter the steady increase in the fraction of households refusing to participate to the survey by raising the planned sample size in order to maintain, or even increase, the achieved sample size. It also devoted considerable efforts to reduce the fraction of non-contacts along two dimensions: number and seasonality. These efforts, as we are about to show, have had important consequences on attrition rates during the period.

#### 4 Attrition and entry

Attrition is a special kind of unit nonresponse arising in longitudinal or rotating surveys, and is defined as failure to obtain data from a sample unit at any wave after it has been selected into the survey. Exogenous or “ignorable” attrition creates no bias and only reduces precision relative to the ideal case of complete data (Little and Rubin 1987). This is no longer true when attrition is endogenous, that is,

systematically related to the response variable of interest. In household surveys, attrition is typically associated with important transitions in a person's life: going to college, finding a new job, marriage or divorce, retirement, death. These events are, at the same time, the main object of study using panel data. Attrition cannot therefore be ignored, for it may lead to invalid inference, even when attrition rates are modest. Similar considerations apply to the new entrants to the survey, for whom information prior to entry is typically not available.

In the case of individuals, it is useful to distinguish between nonmonotone and monotone attrition, depending on whether or not a person who at some point leaves, re-enters the sample at a later date. If attrition and entry (new entry and re-entry) balance each other off, and those who attrite are similar to those who enter the sample, then the sample remains cross-sectionally representative of the target population even when attrition and entry rates are high.<sup>9</sup>

This section describes attrition and new entry in the EPA. We start by describing the general patterns of sample participation and then continue by analyzing how well attrition and new entry may be predicted on the basis of individual characteristics. We conclude by presenting a simple model that tries to capture the main patterns of attrition in the EPA.

#### 4.1 *Patterns of sample participation*

Table 4.1 analyzes the patterns of sample participation of individuals aged 18+ who were first observed during the period from 1987:II to 1997:III. These patterns combine attrition, new entry, and re-entry after nonmonotone attrition.

The top panel shows the 20 most frequent patterns of sample participation (overall, there are  $2^6 - 1 = 63$  possible response patterns), with + denoting response and - nonresponse. Full-time respondents (those who follow the pattern + + + + +) are by far the most frequent case and represent nearly two thirds of the individuals considered. The next two cases by importance are those who only miss the last interview (3.4%) and those who stop participating after the first interview (2.9%). The next four cases consist of those who fail to respond a single interview (2.8% fail the fourth, 2.8% the third, 2.5% the fifth, and 2.3% the second). Other relatively important cases are those who enter the sample after the first interview but stay in continuously thereafter (those who enter in the second and the third interview represent 2.0 and 0.4% respectively), and those who respond once a year, that is, respond only the first and the fifth interview (2.1% of the cases) or only the second and the sixth (0.5% of the cases). The other response patterns are much less frequent.

The bottom panel shows what fraction of those who first responded in wave  $t$  ( $t = 1, \dots, 5$ ) is still in the sample in wave  $t + q$ . Those who were successfully interviewed in wave 1 tend to behave very differently from the others. In fact, the probability of responding in wave  $t + 1$  conditional on starting in wave  $t$  decreases

---

<sup>9</sup> A lesson from the U.S. Panel Study of Income Dynamics (PSID) is that attrition and new entry may contribute to make an initially selected sample more and more representative of the underlying population (see Beckett et al. 1988).

**Table 1.** Patterns of response and persistence of response in the EPA

Most frequent patterns of response					
pattern	%	pattern	%	pattern	%
+++++	64.85	+----+-	2.18	-+----+	0.55
+++++-	3.40	-+++++	2.05	-+-+--+	0.53
+-----	2.90	+--+ +-	1.91	++--++	0.44
+++--++	2.84	+--+-+-	1.90	+++-+--	0.42
++-++++	2.77	-++++-	0.73	--++++	0.40
++++-+	2.50	++++--	0.72	+--+++	0.37
+-----	2.28	-----+	0.59	other	6.22

Persistence of response					
First response in wave	Still present in wave				
	2	3	4	5	6
1	85.7	81.4	77.4	74.0	70.3
2	78.1	61.3	44.4	40.1	—
3	70.4	50.8	35.3	—	—
4	52.9	15.9	—	—	—
5	13.0	—	—	—	—

+ denotes response; —nonresponse. Sample period 1987.II to 1997.III: 1,166,934 individuals aged 18+.

monotonically with *t*. In addition, for those who start in the first wave, the reduction in conditional response rates between the first and the second interview (5%) is much smaller than for the rest of the cases.

In order to illustrate the consequences of nonresponse on labor market outcomes, we compare labor force participation and employment rates in the first interview for two groups: full-time respondents and the other types of sample participants. Table 2 presents the results of such an exercise with data broken down by sex, age group (aged 18–29, 30–39, 40–54, and 55–65), schooling level, and period of the first interview.

While we observe no difference for the central age groups (30–39 and 40–54), we find instead significant differences in the labor force participation and employment rates of the highly educated young (secondary school or college degree) and the elderly with low education (at most primary school). For the former, the labor force participation or employment rates of full-time respondents are lower than the corresponding rates of the other types of sample participants. For the latter the situation is just the opposite, as full-time respondents have higher participation and employment rates than the other types of respondents. Consequently, the EPA underestimates participation of the highly educated young and overestimates participation of the elderly with low education.

**Table 2.** Comparison of labor force participation and employment rates in the first interview of full-time respondents and other respondents

Education	Age Group	Male				Female			
		Participation Diff.	Participation S.e.	Employment Diff.	Employment S.e.	Participation Diff.	Participation S.e.	Employment Diff.	Employment S.e.
Male									
Low	18-29	-3.89	1.28	-0.96	0.69	-1.22	1.82	-0.20	1.06
Low	30-39	0.04	0.71	0.60	0.66	-4.26	2.20	-0.70	0.80
Low	40-54	1.06	0.86	0.56	0.49	-1.25	1.47	-0.03	0.55
Low	55-65	5.84	1.99	1.35	0.69	3.16	1.00	0.87	0.47
High	18-29	-10.82	2.59	-3.71	0.99	-9.99	2.76	-3.08	0.95
High	30-39	0.12	0.55	-0.09	0.84	-3.47	2.12	-1.65	1.71
High	40-54	0.30	0.97	0.11	0.80	-3.15	3.03	-1.24	2.07
High	55-65	2.96	3.97	0.94	2.44	3.33	6.13	1.41	4.35

Diff.: Average of the difference over the period 87.II to 97.III

#### 4.2 Attrition and re-entry

There are several possible definitions of individual and household attrition:

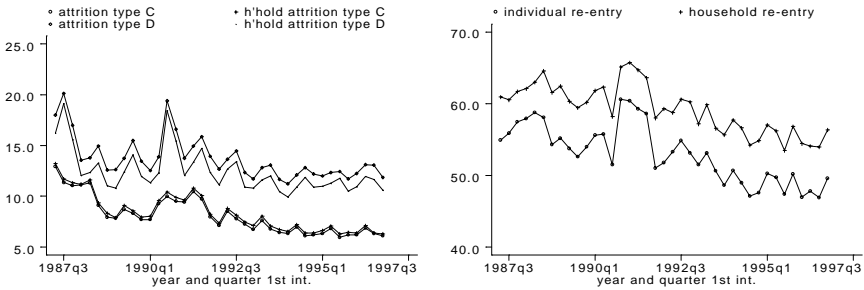
- A. failure to reach the last interview,
- B. failure to respond to any of the six interviews,
- C. failure to being interviewed after one quarter (attrition between successive interviews),
- D. failure to being interviewed after four quarters (attrition over a one-year period).

The first definition is not very informative whereas the second is just the complement of the event of being a full-time respondent. We shall therefore focus on the last two definitions of attrition (C and D), and on re-entry rates given nonresponse in the previous quarter.

The left panel of Fig. 2 presents individual and household attrition rates over the period 1987–1997. No matter what definition is considered, individual attrition rates decrease steadily during the period. The fraction of those who do not respond after one quarter (type C attrition) falls from 11–12 in 1987 to 6% in 1997, whereas the fraction of those who do not respond after one year (type D attrition) falls from 18–20 to 13%. Household attrition rates are of the same order of magnitude as the individual ones. For both individuals and households, moderate seasonality in attrition rates is also present.

The right panel of Fig. 2 shows individual and household re-entry rates at interview  $t$  given nonresponse at interview  $t - 1$ . Re-entry rates are very high (well above 50%) at the beginning of the period, but trend downwards following the parallel decline of attrition rates.





**Fig. 2.** Attrition and re-entry rates by the time of the first interview

Table 3 analyzes the differences in attrition and re-entry rates between successive interviews by key individual characteristics such as sex, age group, schooling level, relationship to the household head and household size.<sup>10</sup>

Because of the stronger attachment, the core of the household (the head, his/her spouse and their children) is less likely to attrite than the other household members (grandparents, other relatives, and other persons living in the household). On the other hand, the largest probability of re-entry is observed for male heads and female spouses, the lowest for daughters with low education and sons-in-law.

Note that, at each age, attrition and re-entry rates differ considerably by education level, and both are higher for the more educated. On the other hand, the general pattern of age dependence is qualitatively the same for each level of education. Attrition rates increase with age up to about age 30, then decline reaching a minimum at about age 50, and rise again afterward. Re-entry rates tends to have the opposite profile. Thus, while attrition and re-entry rates are positively correlated through time, they appear to be negatively correlated over the life-cycle. In particular, attrition rates are high at ages in which important and permanent events occur (typically, marriage and job-related mobility at young ages, and death at older ages), whereas re-entry rates tend to be low at those same ages.

Persons living alone (with no other adult) are more likely to attrite than those living in multi-adult households. In fact, attrition rates mildly decrease with the size of the household. There may be two explanation for this. On the one hand, larger families are less mobile. On the other hand, the proxy respondent method increases the rate of response for larger families.

Although attrition and re-entry rates tend to be positively correlated, there are some important exceptions. In particular, children and other relatives are more likely to attrite and less likely to re-enter the survey after missing an interview. The same is also true for people in single-adult households.

<sup>10</sup> Although not reported in the table, important regional differences have also been detected. Canary Islands and Madrid show attrition rates that are almost twice as high as those of La Rioja or Galicia. In general, the regions with the highest attrition rates also show very high re-entry rates.

**Table 3.** Individual attrition and re-entry between successive interviews by individual characteristics.<sup>a</sup>

		Level of schooling								Total
		None/Primary		Secondary		College		All levels		
		Male	Female	Male	Female	Male	Female	Male	Female	
<b>By age group</b>										
18–24	C	7.6	9.4	6.5	6.9	7.2	7.5	7.1	8.0	7.5
	R	42.9	37.1	53.8	52.4	60.9	56.7	48.7	45.8	47.2
25–29	C	10.0	10.5	10.1	10.8	10.8	11.6	10.2	10.8	10.5
	R	41.0	46.0	43.9	46.0	43.4	43.4	42.3	45.3	43.8
30–34	C	9.0	8.6	10.6	10.2	14.0	13.2	10.1	9.8	9.9
	R	50.4	55.5	51.7	53.8	50.2	52.1	50.7	54.3	52.5
35–39	C	7.9	7.6	9.7	9.5	12.5	12.4	8.9	8.5	8.7
	R	57.1	59.4	56.3	59.6	57.8	59.9	57.1	59.6	58.3
40–44	C	6.9	6.4	8.8	8.9	12.2	11.3	7.8	7.1	7.5
	R	60.0	61.6	60.8	59.1	60.8	63.2	60.3	61.5	60.9
45–49	C	6.1	5.8	7.5	8.5	10.2	10.2	6.7	6.2	6.4
	R	58.8	60.6	61.6	62.0	63.6	64.9	60.0	61.2	60.6
50–54	C	5.8	5.9	7.4	9.2	9.6	10.0	6.2	6.2	6.2
	R	57.9	60.8	59.4	55.6	63.5	62.4	58.8	60.6	59.7
55–59	C	6.0	6.5	9.0	10.0	9.3	10.5	6.4	6.7	6.6
	R	59.5	61.4	58.0	61.7	63.2	62.2	59.7	61.4	60.6
60–64	C	6.7	7.4	8.4	11.2	9.7	11.1	7.0	7.6	7.3
	R	60.0	61.7	61.8	63.3	61.1	64.1	60.2	61.8	61.1
65–69	C	7.5	8.3	10.3	11.9	11.1	12.4	7.8	8.4	8.1
	R	59.9	61.6	57.6	61.4	66.4	61.4	60.3	61.6	61.0
70+	C	8.6	9.6	11.1	11.1	12.2	13.0	8.9	9.7	9.3
	R	48.9	47.8	55.3	55.3	57.8	59.9	49.7	48.2	48.8
<b>By relationship with the head of the household</b>										
Head	C	7.4	10.1	10.4	14.7	12.2	16.5	8.2	10.8	8.7
	R	58.0	53.4	57.4	48.7	59.4	52.9	58.1	53.0	56.9
Spouse	C	9.5	7.0	13.3	10.5	14.6	12.0	11.1	7.6	7.7
	R	45.0	61.5	46.0	59.3	47.4	60.7	45.7	61.2	60.9
Child	C	6.8	7.4	6.7	6.6	8.2	8.1	6.9	7.2	7.1
	R	42.6	39.1	50.8	50.6	51.2	51.0	46.7	45.7	46.2
Child	C	7.2	7.8	8.2	9.3	10.0	10.4	7.5	8.1	7.7
In law	R	33.6	28.5	33.9	37.4	25.7	34.7	33.1	30.3	32.1
Other	C	10.8	10.9	11.5	12.8	15.3	16.4	11.1	11.3	11.2
	R	33.7	36.5	31.7	35.7	36.3	37.8	33.6	36.5	35.5
<b>By household size (number of members aged 16+)</b>										
Single	C	14.9	13.4	23.5	19.0	22.5	19.8	16.6	14.1	14.8
	R	45.8	52.4	37.2	46.4	45.4	51.6	44.6	51.9	49.7
Two	C	9.4	9.3	11.7	11.8	14.0	14.0	10.2	10.0	10.1
	R	57.5	58.5	56.0	56.4	57.7	57.5	57.3	58.1	57.7
Three	C	6.8	7.0	8.1	8.3	10.3	10.3	7.3	7.4	7.4
	R	57.5	58.5	56.0	56.4	57.7	57.5	57.3	58.1	57.7
Four	C	6.0	6.2	6.9	6.9	8.4	8.5	6.4	6.5	6.5
	R	51.0	51.7	53.6	53.3	56.9	54.8	52.3	52.4	52.4
Five	C	5.8	6.2	6.0	6.2	7.6	7.6	6.1	6.4	6.2
	R	44.3	43.5	48.6	48.5	54.5	49.9	46.6	45.4	46.0
All	C	7.5	7.9	8.4	8.6	10.9	10.9	8.0	8.3	8.1
	R	52.9	54.1	53.1	53.1	56.0	54.8	53.4	54.0	53.7

<sup>a</sup> Sample selection: Aged 18+, first interview between 1987:II and 1996:III. C: Type-C attrition; R: Re-entry.

**Table 4.** Unconditional nonresponse rates and conditional attrition rates

	Quarter of interview					
	1	2	3	4	5	6
Unconditional nonresponse rates	7.7	15.8	14.5	13.3	12.0	17.0
Conditional attrition rates given						
$R_{t-1} = -$	-	33.6	59.5	54.3	46.8	46.2
$R_{t-1} = +$	-	14.3	6.1	6.4	6.6	13.0
$R_{t-1} = -, R_{t-2} = -$	-	-	55.8	66.7	58.0	60.3
$R_{t-1} = -, R_{t-2} = +$	-	-	60.3	31.6	30.7	30.8
$R_{t-1} = +, R_{t-2} = -$	-	-	21.9	14.8	15.0	45.6
$R_{t-1} = +, R_{t-2} = +$	-	-	5.0	5.7	5.9	10.1

+ Denotes response and - nonresponse. Sample selection: aged 18+, first interview between 1987:II and 1997:III.

### 4.3 A simple model of attrition and re-entry

Table 4 presents individual unconditional nonresponse rates and conditional (one-period) attrition rates by interview number. As shown in the top panel of the table, unconditional nonresponse rates increase with the interview number from 7.6% in the first interview to 17% in the last. On the other hand, conditional attrition rates (bottom panel) given nonresponse in the previous quarter start at 33.6% in the second wave and increase somewhat in the third wave to decrease afterward. Conditional on response in the previous quarter, attrition rates start instead at 14.3% in the second wave and decrease slightly to 6.1% in the third wave. When conditioning on behavior in two consecutive quarters, the dominant effect seems to be whether or not the person was a respondent in the previous quarter, with a smaller role for responding two quarters before. Overall, these results suggest that a simple Markov chain model may provide a good representation of the patterns of sample participation in the EPA.

Thus consider a model of individual attrition and re-entry based on the following set of assumptions:

1. attrition is the only source of nonresponse;
2. all people are successfully interviewed in the first wave ( $t = 1$ );
3. between any two waves  $t$  and  $t + 1$ , sample participants at time  $t$  attrite with probability  $\phi$  and non participants at time  $t$  re-enter with probability  $\eta$ .

The second assumption is really unnecessary, but it makes things a little easier. If  $\pi_t$  denotes the probability of sample participation at wave  $t$ , then our assumptions imply that

$$\begin{pmatrix} \pi_t \\ 1 - \pi_t \end{pmatrix} = \Psi \begin{pmatrix} \pi_{t-1} \\ 1 - \pi_{t-1} \end{pmatrix}, \quad t > 1,$$

where  $\pi_1 = 1$  (because of Assumption 2) and

$$\Psi = \begin{bmatrix} 1 - \phi & \eta \\ \phi & 1 - \eta \end{bmatrix}, \quad (1)$$

with  $0 \leq \phi, \eta \leq 1$ . Hence, the probability of sample participation evolves through time according to

$$\begin{pmatrix} \pi_t \\ 1 - \pi_t \end{pmatrix} = \Psi^{t-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad t > 1.$$

A little algebra shows that

$$\pi_t = \eta + \psi\pi_{t-1},$$

where  $\psi = 1 - \phi - \eta$ . Repeated backward substitution then gives

$$\pi_t = \eta(1 + \psi + \psi^2 + \dots + \psi^{t-2}) + \psi^{t-1}.$$

In the special case when  $\eta = 0$  (monotone attrition), we have that  $\pi_t = (1 - \phi)\pi_{t-1} = \dots = (1 - \phi)^{t-1}$  and so  $\pi_{t-1} > \pi_t$  for all  $t$ . If  $\eta > 0$  (nonmonotone attrition), then the steady-state probability of sample participation is

$$\pi^* = \lim_{t \rightarrow \infty} \pi_t = \frac{\eta}{1 - \psi} = \frac{\eta}{\phi + \eta}.$$

In what follows, we shall use  $\pi^*$  as a summary measure in order to make comparisons across different socio-demographic groups.

Both the attrition rate  $\phi$  and the re-entry rate  $\eta$  are assumed to depend on a covariate vector  $X$  which includes sex, age, education level, marital status, household size, number of children, other individual characteristics and household background variables, controls for the interview number and quarter of the interview and period of the first interview effects. The latter may be viewed as proxies for the effort by INE to control the sample dynamics. To model the dependence of  $\phi$  and  $\eta$  on  $X$ , we consider the conditional logit specifications

$$\ln \frac{\phi(X)}{1 - \phi(X)} = \alpha + X'\beta, \quad \ln \frac{\eta(X)}{1 - \eta(X)} = \gamma + X'\delta,$$

where  $\alpha, \beta, \gamma$  and  $\delta$  are unknown parameters to be estimated.

We fit each model using two alternative samples. The first is restricted to individuals who respond to the first interview, whereas the second is unrestricted (that is, it considers entrants at interview 2,3,...). In both cases, we take a random subset (1/10) of individuals from the pooled EPA waves.

Although most coefficients are statistically significant, mainly because of the very large sample size, the models as a whole capture only a small fraction of the observed behavior (the pseudo  $R^2$  does not exceed 5% for the models of attrition and 8% for the models of re-entry). Instead of presenting the estimated coefficients, Table 5 reports the estimated probabilities for the baseline case, along with the ones obtained by varying one characteristic at the time with respect to the baseline. The

baseline is a male household head aged 30–39, married, with at least one child aged 0–5, interviewed in the second wave, with secondary education, full-time employed, with a permanent contract and previous work experience. For the first sample, the implied steady-state probability of sample participation is about 75%.

Note first that the baseline probability of attrition is higher in the first sample, namely the one restricted to first interview respondents, whereas the probability of re-entry is lower. Second, in all models, attrition and re-entry probabilities decrease with the interview number. Third, attrition is much more important in the third quarter and, as a result, re-entry rates are higher in the fourth quarter. In all cases, the null hypothesis that the interview number or the quarter of the interview have no effect on attrition or re-entry probabilities is strongly rejected.

Turning to individual and household characteristics, people who are not married have lower probability to attrite. Attrition probabilities reach a minimum around age 50, whereas re-entry probabilities peak around age 60. While both types of attrition increase with the level of schooling, re-entry is hardly affected by education. The probability of attrition is lower for the core of the household and higher for the other members. The probability of re-entry is much lower for adult children (16+) and other members of the household than for the head and his/her spouse.

The number of household members who are employed has little effect on attrition and re-entry rates. The presence of small children (aged less than 16) significantly reduces attrition probabilities but has no significant effect on re-entry. Attrition probabilities decrease with the size of the family, whereas re-entry probabilities peak for households of size 3.

Labor market status and the other variables which describe employment characteristics, although significant in some cases, have little effect on attrition and re-entry probabilities. On the contrary, being at school or being a family business helper sharply reduces attrition and increases re-entry probabilities, whereas the willingness to migrate both increases attrition probabilities and reduces re-entry probabilities.

There are notable differences by region. Attrition is lower in the North, Center and East, than in the South and Madrid. However, the regions that show higher probability of attrition also display higher probability of re-entry.

Our final comments regard the Wald tests reported at the bottom of the table. Stability of attrition and re-entry rates over time and across interview number and quarter is always strongly rejected. In particular, attrition rates mildly decrease with the interview number and increase significantly during the third quarter. Our regression results also show a long-term decreasing trend in attrition rates, which does not appear to be the result of compositional effects.

## 5 Labor force transitions

This section studies transitions between labor force states. We address three issues. First, we compare annual with quarterly transition rates without taking attrition into account, and analyze the resulting pattern of transitions by age and sex. Second, we construct parametric and nonparametric bounds for population transition

probabilities in order to provide a measure of the degree of uncertainty caused by the presence of sample attrition. Third, we perform two semiparametric tests of the impact of sample attrition: the first compares transitions of full-time respondents with those of other sample participants, the second is a regression-based test on grouped data.

### 5.1 A comparison of annual and quarterly transitions

Table 6 presents quarterly (Q) and annual (A) transition rates between four labor force states, namely full-time employment (FT, 35+ weekly hours of work), part-time employment (PT, less than 35 weekly hours), unemployment (U), and out of the labor force or inactivity (O). Tabulations are broken down by sex, age group (aged less than 30, 30–49, 50–64), and level of education: lower (at most primary) and higher (secondary or university). As one would expect, retention rates in the same state are always lower if measured on annual rather than quarterly basis, whereas the opposite is true for transitions from one state to another.

A direct comparison of quarterly and annual transition rates is not very meaningful, however, because of the different period of reference. To overcome this difficulty, we instead compare annual transition rates with annualized quarterly transition rates. Thus, consider the following relationship between state probabilities one quarter apart

$$\Pi_{t+1} = \Lambda_t^\top \Pi_t$$

where  $\Pi_t$  is the vector of state probabilities in period  $t$ ,  $\Lambda_t = [\lambda_{t,ij}]$  is the quarterly transition probability matrix, and  $\lambda_{t,ij}$  is the quarterly transition probability from state  $i$  in quarter  $t$  to state  $j$  in quarter  $t+1$  ( $i, j = \text{FT, PT, U, O}$ ). The corresponding relationship between state probabilities one year apart is

$$\Pi_{t+4} = \tilde{\Lambda}_t^\top \Pi_t.$$

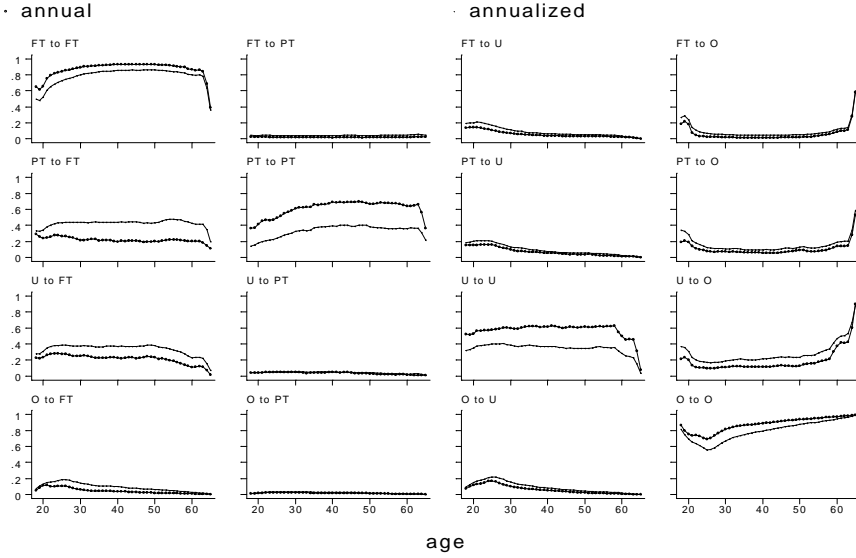
If there were no attrition then  $\tilde{\Lambda}_t = \Lambda_t^*$ , with  $\Lambda_t^* = \prod_{k=0}^3 \Lambda_{t+k}$ .

Are observed annual and annualized transition matrices mutually consistent? Clearly not, since observed annual transition matrices tend to overestimate retention rates and underestimate transition rates due to attrition of the more mobile workers. As an example, Fig. 3 compares the annual transition matrix  $\tilde{\Lambda}_t = [\tilde{\lambda}_{t,ij}]$  with its annualized version  $\Lambda_t^* = [\lambda_{t,ij}^*]$ , both as a function of age. In general, annualized transitions overestimate retention rates and underestimate transition rates between states. However, while for retention rates the difference is always significant, for transition rates the results are more mixed. In particular, annual transitions appear to underestimate considerably exit from unemployment back into full-time work. On the other hand, the difference between the two sets of estimates is small in the case of transition rates to part-time employment or unemployment, which in general are not very important.

**Table 5.** Estimated probabilities from logit models of attrition and re-entry

	Attrition		Re-entry		Sample participation	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
Baseline	19.3	17.5	56.4	60.1	74.5	77.4
Interview 3	17.5	15.6	48.1	51.3	73.3	76.7
Interview 4	16.9	14.7	41.3	43.7	71.0	74.8
Interview 5	16.5	13.2	35.0	35.2	68.0	72.7
Interview 6	15.9	13.7	–	–	–	–
2nd quarter	19.3*	17.4*	49.0	52.6	71.7	75.1
3rd quarter	31.6	29.2	46.0	49.7	59.3	63.0
4th quarter	24.6	23.0	65.2	68.0	72.6	74.7
Female	19.3*	17.4*	55.9*	59.4*	74.3	77.3
Single	16.4	16.3	56.2*	58.5*	77.4	78.2
Age 25	23.4	21.6	50.4	54.1	68.3	71.5
Age 45	16.6	15.2	61.4	63.8	78.7	80.8
Age 55	14.4	13.5	62.6	65.6	81.3	82.9
Age 65	14.8	13.8	64.1	66.6	81.2	82.8
Age 75	15.1	14.4	60.8	62.1	80.1	81.2
Age 82	17.3	16.1	49.0	44.4	73.9	73.4
None/primary educ.	16.6	14.3	56.3*	59.1*	77.2	80.5
College	22.9	20.3	57.4*	58.4*	71.5	74.2
Spouse of the head	18.3	17.2*	60.8	62.4	76.9	78.4
Son of the head	18.1	15.6	49.6	51.5	73.3	76.8
Other members	29.0	24.4	40.3	43.3	58.2	64.0
number of employed in the hh	19.2*	17.5*	55.8*	59.8*	74.4	77.4
No children 0–5	21.6	20.4	57.5*	60.9*	72.7	74.9
Children 6–10	17.3	15.6	57.1*	60.4*	76.7	79.5
Children 11–15	17.4	15.8	57.3*	60.6*	76.7	79.3
Household size of 1	29.6	27.6	50.9	55.9	63.2	66.9
Household size of 3	14.5	13.0	53.8	59.1*	78.8	82.0
Household size of 4	12.8	11.3	52.8	55.9	80.5	83.2
Household size of 5	11.9	10.5	47.1	51.8	79.8	83.1
Unemployed	19.2	17.6*	56.1	62.4*	74.5	78.0
Out of the labor force	19.0	17.6*	56.8	62.9*	74.9	78.1
Temporary contract	17.9	16.9	53.7	57.5	75.0	77.3
Civil servant	20.4	18.4	53.2	62.1	72.3	77.1
No full time work	18.7*	17.3*	53.2	61.6	74.0	78.1
Does on the job search	19.9*	17.3*	54.6*	57.0*	73.3	76.7
No previous experience	19.4*	17.9*	58.0	61.7	74.9	77.5
Family business helper	16.7	14.9	59.1*	63.3*	78.0	80.9
Studying	17.5	15.4	65.0	68.5	78.8	81.6
Willingness to migrate	20.4	19.5	53.0	58.0*	72.2	74.8
Does domestic work	17.1	16.1	59.9	59.3*	77.8	78.6
South	16.9	15.8	57.5*	60.5*	77.3	79.3
Centre	15.5	14.6	56.6*	61.1*	78.5	80.7
East	15.4	14.5	54.5	56.9	78.0	79.7
North	14.7	13.8	52.2	56.6	78.0	80.4
Number of observations	449595	427164	50367	45415		
Wald $W_0$ : all coefficients	10075.2	9467.7	679.5	4709.2		
Wald $W_1$ : int. and quarter dummies	2696.8	2642.8	2068.2	2149.2		
Wald $W_2$ : group of entry dummies	2018.1	1853.3	469.4	425.2		
Wald $W_3$ : joint $W_1$ and $W_2$	4881.1	4527.8	2517.7	2568.6		
Pseudo $R^2$	0.040	0.041	0.070	0.078		

Baseline: Interviewed in 2 (not interviewed in 3 for re-entry), male, age 35, secondary studies, head, married, at least a child 0–5, living in Madrid with another adult (hh size 2), full-time employed with an indefinite contract and with previous work experience, and the rest of the variables taking a negative response. Other variables: period of the first interview dummies (39), age, age<sup>2</sup>, age<sup>3</sup>. Sample selection criteria: one tenth of the relevant sample (1987:II to 1997:IV). Sample 1: All individuals. Sample 2: Only individuals who responded to the first interview. Note: starred probabilities denote insignificant (at 5%) differences with respect to the baseline case.



**Fig. 3.** Annual and annualized transition rates by age. All the figures have been smoothed out using a centered 3-year moving average. *FT*: Full-time employment; *PT*: Part-time employment; *U*: Unemployment; *O*: Out of the labor force

### 5.2 Bounds on transition probabilities

In this section we evaluate the degree of uncertainty caused by sample attrition by constructing bounds within which the population transition probabilities must lie.

Consider the situation arising in a transition model with 2 periods and 4 states (full-time employment, part-time employment, unemployment, and out of the labor force). The estimated transition rate between state  $i$  and state  $j$ , denoted by  $l_{ij}$ , is given by

$$l_{ij} = \frac{m_{ij}}{m_i},$$

where  $m_{ij}$  is the observed flow from state  $i$  in period 1 to state  $j$  in period 2, and  $m_i = \sum_{j=1}^4 m_{ij}$  is the number of people in state  $i$  in period 1 whose records can be matched across the two periods. If the sample size is sufficiently large, then the true population transition probability  $\lambda_{ij}$  must necessarily lie in the interval

$$[l_{ij}^*, \bar{l}_{ij}^*] = \left[ \frac{m_{ij}}{m_i + u_i}, \frac{m_{ij} + u_i}{m_i + u_i} \right], \quad (2)$$

where  $u_i$  denotes the number of unmatched records from state  $i$  in period 1. The lower bound is the transition rate if all the unmatched remain in state  $i$ , while the upper bound is the transition rate if they all move to state  $j$ . The interval (2) may equivalently be expressed as

$$\left[ \frac{m_{ij}}{m_i} (1 - f_i), \frac{m_{ij}}{m_i} (1 - f_i) + f_i \right],$$



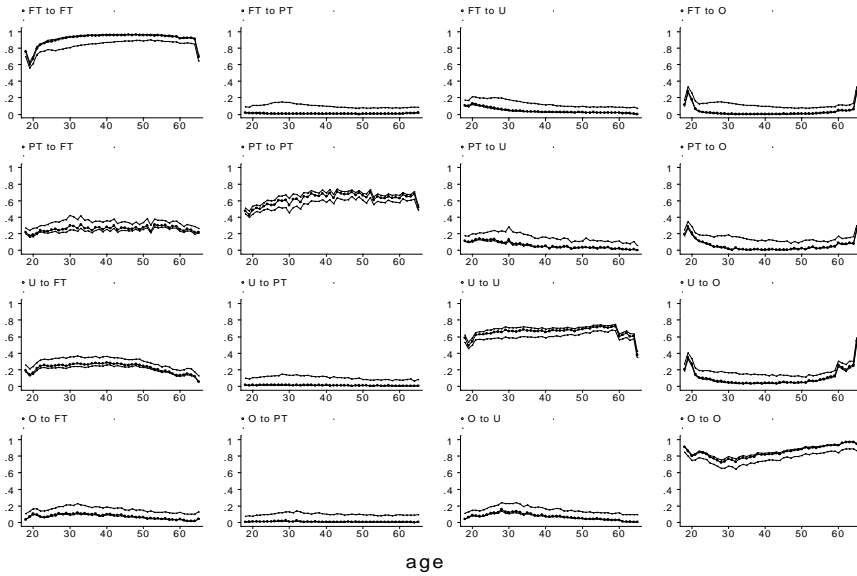


Fig. 4. Transition probabilities and bounds by age, Men

where  $f_i = u_i / (m_i + u_i)$  is the attrition rate among those in state  $i$  in period 1. This form of the bounds coincides with expression (2) in Horowitz and Manski (1998) when only the outcome variable is censored and its support is the unit interval. It shows that the width of the interval (2) is equal to the attrition rate from state  $i$ . The latter may therefore be viewed as a measure of how “vague” is the information about  $\lambda_{ij}$  contained in the matched data. The lower is the attrition rate, the tighter are the bounds and, therefore, the more informative the matched data are about the population transition probabilities.

Following Peracchi and Welch (1995), Fig. 4 presents the observed transition rates for the selected sample and the estimated bounds age and labor force status in the first period for men.<sup>11</sup> It should be stressed that, for a majority of transitions, the bounds implied by (2) are rather tight, although their width (the estimated attrition rate) varies considerably by age and sex. Transitions from full-time to part-time employment are a good example of variations of the bound by age, whereas retention rates in inactivity are a good example of differences in the shape of the bounds by gender. Although the figure does not report classical confidence intervals at the conventional 95 % level, it turns out that they are always contained in our bounds. This shows that the uncertainty caused by sample attrition is always larger than the one caused by sampling variability.

Because of the “curse of dimensionality” problem, nonparametric bounds like the ones presented in the above figure only allow to control for a very limited number of individual or household characteristics. To avoid this limitation, we also estimate parametric models for the transition probabilities  $\lambda_{ij}$  and the attrition probabilities

<sup>11</sup> The corresponding figure for women is not reported for brevity, but is available from the authors upon request.

**Table 6.** Transitions probabilities by age group and education level<sup>a</sup>

sex	age	type	to/from	Lower education				Higher education			
				FT	PT	U	O	FT	PT	U	O
M	16–29	Q	FT	88.74	29.30	21.73	5.85	91.01	19.40	17.14	3.03
M	16–29	Q	PT	0.95	54.15	1.42	0.69	1.33	64.96	2.08	0.71
M	16–29	Q	U	7.84	10.41	70.42	7.81	4.83	6.48	69.16	5.03
M	16–29	Q	O	2.47	6.14	6.43	85.65	2.83	9.16	11.62	91.23
M	16–29	A	FT	80.40	39.37	34.30	20.84	84.08	27.88	31.81	8.85
M	16–29	A	PT	1.16	33.20	2.19	1.71	1.48	44.16	3.70	1.67
M	16–29	A	U	11.77	15.76	52.08	17.15	8.43	11.80	47.54	9.13
M	16–29	A	O	6.67	11.66	11.43	60.30	6.01	16.16	16.95	80.35
M	30–49	Q	FT	95.69	36.09	22.62	6.52	97.27	20.87	16.87	9.78
M	30–49	Q	PT	0.55	54.84	1.18	0.54	1.27	76.98	2.14	1.28
M	30–49	Q	U	3.21	7.16	72.82	6.89	1.17	1.64	76.81	12.93
M	30–49	Q	O	0.55	1.91	3.38	86.05	0.29	0.51	4.17	76.00
M	30–49	A	FT	93.07	45.22	37.25	10.10	95.45	27.79	34.89	18.11
M	30–49	A	PT	0.66	41.77	1.64	1.13	1.66	68.54	3.65	2.71
M	30–49	A	U	5.24	10.17	56.53	8.95	2.35	2.84	56.82	15.62
M	30–49	A	O	1.04	2.84	4.58	79.81	0.53	0.83	4.63	63.56
M	50–69	Q	FT	94.66	27.72	13.96	0.65	96.41	19.09	8.21	0.81
M	50–69	Q	PT	0.96	63.14	0.86	0.19	1.56	78.27	0.83	0.23
M	50–69	Q	U	1.93	2.35	74.55	0.79	0.48	0.36	81.06	0.78
M	50–69	Q	O	2.45	6.80	10.63	98.37	1.54	2.28	9.90	98.17
M	50–69	A	FT	87.68	30.31	20.45	1.09	91.68	24.29	23.72	1.76
M	50–69	A	PT	1.16	51.01	1.31	0.39	1.89	67.80	1.19	0.61
M	50–69	A	U	3.34	2.89	57.14	0.89	1.20	0.95	59.03	1.02
M	50–69	A	O	7.82	15.79	21.11	97.63	5.24	6.96	16.07	96.60
F	16–29	Q	FT	84.74	16.43	10.42	2.64	88.29	14.81	10.09	1.86
F	16–29	Q	PT	3.00	68.67	2.81	0.92	2.77	70.65	3.12	0.82
F	16–29	Q	U	8.67	9.41	77.84	5.77	6.55	8.41	76.37	5.76
F	16–29	Q	O	3.59	5.50	8.93	90.67	2.39	6.13	10.42	91.55
F	16–29	A	FT	76.38	22.75	18.39	6.21	81.00	22.08	20.67	4.98
F	16–29	A	PT	3.84	50.71	5.66	2.24	3.56	54.47	6.19	1.90
F	16–29	A	U	13.93	16.45	64.06	11.63	11.56	13.79	60.06	10.81
F	16–29	A	O	5.85	10.08	11.89	79.91	3.89	9.67	13.08	82.31
F	30–49	Q	FT	90.31	12.63	6.67	1.54	94.16	14.07	7.23	1.95
F	30–49	Q	PT	3.23	78.46	2.84	0.83	3.30	82.09	2.60	1.07
F	30–49	Q	U	3.18	3.58	77.47	2.54	1.74	2.33	79.88	5.87
F	30–49	Q	O	3.28	5.33	13.03	95.09	0.81	1.51	10.29	91.11
F	30–49	A	FT	85.82	16.30	10.08	2.80	90.65	17.97	14.24	4.56
F	30–49	A	PT	3.98	66.53	6.18	2.01	4.37	74.27	5.21	2.51
F	30–49	A	U	5.08	6.69	64.09	4.34	3.28	4.62	65.51	8.72
F	30–49	A	O	5.11	10.47	19.65	90.85	1.70	3.13	15.04	84.22
F	50–69	Q	FT	90.33	13.55	6.01	0.49	93.38	13.71	4.79	0.44
F	50–69	Q	PT	4.20	78.34	1.92	0.32	4.45	84.22	2.12	0.36
F	50–69	Q	U	1.12	1.14	71.98	0.41	0.36	0.41	78.64	0.69
F	50–69	Q	O	4.35	6.98	20.09	98.78	1.80	1.66	14.46	98.51
F	50–69	A	FT	84.35	15.37	6.47	0.83	87.66	16.41	8.06	1.06
F	50–69	A	PT	5.13	66.45	3.81	0.74	5.81	77.09	3.88	0.82
F	50–69	A	U	1.49	2.17	57.26	0.64	0.79	0.78	65.07	1.25
F	50–69	A	O	9.03	16.01	32.47	97.79	5.74	5.73	22.99	96.87

<sup>a</sup>Sample period: 1987:III to 1997:IV. FT: Full-time employment; PT: Part-time employment; U: Unemployment; O: Out of the labor force.

$\phi_i$  that allow for a broad set of covariates, including calendar time, interview number, sex, age, education level, marital status, household size, number of children, and other individual characteristics and household background variables.

Denoting by  $\hat{\phi}_i(x)$  a parametric estimate of attrition probability for an individual in state  $i$  with covariate vector  $x$  (based on the model in Sect. 4.3) and by  $\hat{\lambda}_{ij}(x)$  the corresponding parametric estimate of transition probability from state  $i$  to state  $j$  based on the matched data, then a parametric estimate of the bound on population transition probabilities is

$$\left[ \hat{\lambda}_{ij}(x) (1 - \hat{\phi}_i(x)), \hat{\lambda}_{ij}(x) (1 - \hat{\phi}_i(x)) + \hat{\phi}_i(x) \right].$$

Figure 5 illustrates the results obtained for retirement rates, defined as transitions from labor force participation (either employment or unemployment) to inactivity.<sup>12</sup> Estimates refer to the second quarter and are presented separately by sex, schooling (primary vs. university level) and year of the first interview (1987, 1990, 1993 and 1996).

Retirement rates increase with age, with a clear peak at age 65, and are higher for women and less educated workers. On the other hand, the width of the bounds shows exactly the opposite pattern, for it decreases with age, and is lower for women and less educated workers.

### 5.3 Testing selection in transition probabilities

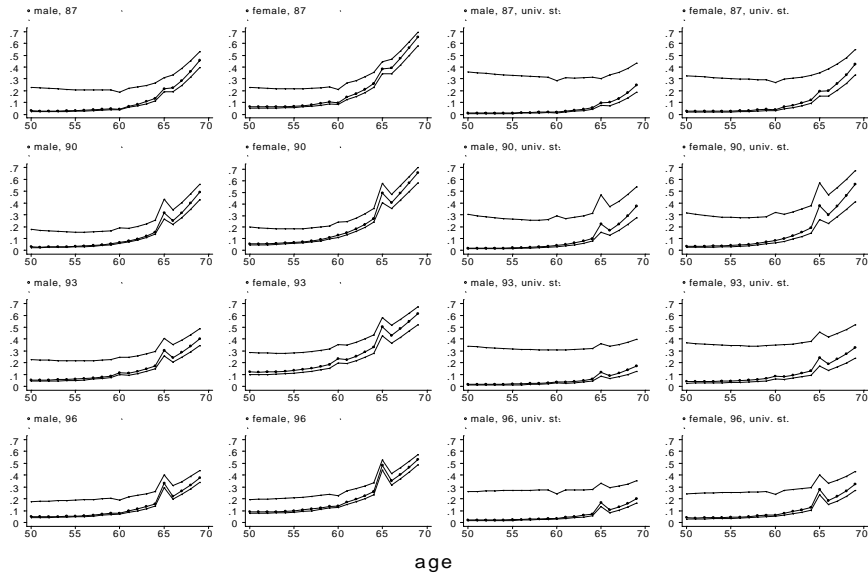
In this section we consider more formal tests of the null hypothesis that attrition does not cause biases in estimates of transition based on matched data. The basic idea is to compare transitions rates of full-time respondent, that is, individuals who are in sample for all six interviews, with transitions rates of those who are in the sample for less than six periods. Statistically significant differences between the two sets of estimates are taken as evidence against the null hypothesis.

Thus, let  $Y_{ij}$  denote the indicator of whether or not a person moves from state  $i$  at time  $t$  to state  $j$  at time  $t + 1$  ( $Y_{ij} = 1$  for movers), and let  $D$  denote the indicator of whether or not a person is a full-time respondent ( $D = 1$ ). Attrition gives rise to a bias if

$$\lambda_{ij}^{(0)}(x) = E(Y_{ij} | X = x, D = 0) \neq E(Y_{ij} | X = x, D = 1) = \lambda_{ij}^{(1)}(x),$$

where  $X$  is a set of conditioning variables. When  $X$  is a discrete random vector, we can partition the sample on the basis of the observed values  $x$  of  $X$  and carry out the test nonparametrically by just looking at the differences between the observed transition rates  $l_{ij}^{(0)}(x)$  and  $l_{ij}^{(1)}(x)$  for the two groups. Critical values for the test

<sup>12</sup> The sample used to estimate the transition and nonresponse models has been restricted to individuals aged 50–69. The variables considered are the interview number, sex, marital status, a polynomial in age, the relationship with the reference person, the number of employed people within the household (the individual excluded), the number of children below 0–5, 6–10, 11–15, the number of adult people, and the region of residence. The estimation results are available from the authors upon request.



**Fig. 5.** Parametric transitions out of the labor force and bounds by age, sex and year of the first interview. Base case: married, lower education, head (spouse if female), no other person employed in the family, living in the South, in a 2-person household (no children)

are based on the fact that the difference  $l_{ij}^{(0)}(x) - l_{ij}^{(1)}(x)$  divided by its estimated standard error is distributed as Student's  $t$  under the null hypothesis of equality.

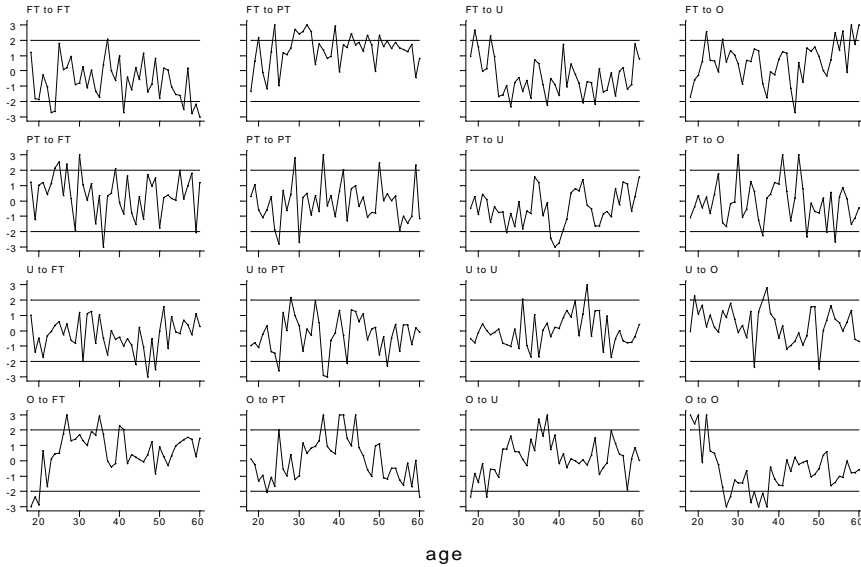
Figure 6 reports the value of the test statistic for cells defined by age group and labor force status in the first period for men.<sup>13</sup> In most cases, the test statistic is less than 2 in absolute value, indicating that attrition and re-entry are generally not a source of concern for the estimation of transitions probabilities. In some special cases, however, conclusions are a little less reassuring. In particular, equality is rejected for transition of young people (aged 25–30) from inactivity to employment and for transition of people aged 50+ from employment to inactivity.

To further analyze the evidence of attrition bias, consider the following decomposition of population transition probabilities

$$\lambda_{ij} = \sum_k E(Y_{ij}|D = k) \Pr\{D = k\},$$

where  $D$  is an indicator for the type of sample participation and the dependence on a vector  $X$  of covariates has been omitted for simplicity. We distinguish between participation in both periods ( $D = 1$ ), participation in the first but not the second period ( $D = 2$ ), participation in the second but not the first period ( $D = 3$ ), and participation in neither period ( $D = 4$ ). Since  $\Pr\{D = 1\} > 0$ , following Peracchi

<sup>13</sup> The corresponding figure for women is not reported for brevity, but is available from the authors upon request.



**Fig. 6.** Test statistics for the equality of quarterly transitions probabilities of full-time respondents and other sample participants, Men. (Note: Age 60 groups individuals aged 60–64.)

and Welch (1995), we have

$$E(Y_{ij} | D = 1) = \frac{\lambda_{ij}}{\Pr\{D = 1\}} - \sum_{k>1} E(Y_{ij} | D = k) \frac{\Pr\{D = k\}}{\Pr\{D = 1\}}.$$

The hypothesis of absence of attrition bias is then equivalent to the hypothesis that the coefficients in the regression model

$$E(Y_{ij} | D = 1) = \beta_1 \frac{1}{\Pr\{D = 1\}} + \sum_{k>1} \beta_k \frac{\Pr\{D = k\}}{\Pr\{D = 1\}} \tag{3}$$

are all equal, except for the sign. To carry out the test, the data have been grouped by age and age. For each type of transition and each age-sex combination, a regression is estimated exploiting the variation either by province or by quarter of the first interview. In all cases, we carry out weighted and unweighted regressions. In Table 7 we report the average significance levels of the Wald test for equality. In most cases, we cannot reject the null hypothesis that the coefficient of the regression (3) are equal. The only exceptions occur with transition from part-time employment and, in some cases, from unemployment.

## 6 Conclusions

In this paper we consider the implications of sample attrition for the study of labor force dynamics using data from the Spanish labor force survey. This is an

important issue because, in order to use the EPA to predict labor market outcomes, one would like to be reasonably confident about the quality of this source. Our main conclusions may be summarized as follows.

First, survey nonresponse cannot be neglected. Despite the efforts by the INE, nonresponse rates have been steadily increasing and are likely to further increase in the future as the composition of the population shifts towards groups, such as the more educated, that are characterized by higher nonresponse rates.

Second, we showed that, as a result of differential nonresponse rates by education and age group, the EPA tends to underestimate employment and participation rates of young people with higher education, and to overestimate employment and participation rates of the elderly with lower education.

Third, the patterns of attrition and re-entry into the survey may be summarized by a simple Markov-chain model where attrition and re-entry probabilities depend both on observable individual characteristics and other variables, including a time trend which may capture the effort by the INE to control the sample dynamics. However, stability of attrition and re-entry probabilities over time and across quarter and interview number is strongly rejected by the data.

Fourth, it makes a difference whether transition rates are estimated on an annual basis or by annualizing quarterly transitions. For example, annual transitions overestimate the probability that older workers stay in the same state and underestimate the probability that they move from one state to another.

Finally, similar to what has been found for other surveys, we find little evidence that attrition causes important biases in quarterly transition probabilities estimated from the matched data. The main exceptions are transitions of young people from inactivity to employment and transitions of those aged 50+ from employment to out of the labor force.

## Appendix: Matching the EPA

Starting with the second quarter of 1987, INE releases public-use longitudinal matched EPA files, called *Estadística de Flujos* (EF), obtained by exploiting the rotating nature of the survey. They contain fewer variables than the regular EPA, but allow sample persons to be followed for their duration of stay in the survey, namely up to 6 quarters. These data have several limitations which make them not well suited for analyzing labor force transitions:

1. The household identifier and the residence are dropped from the tape.
2. Age is grouped in coarse (5-year) brackets.
3. A few important variables such as the sector and the occupation of employment are recoded into fewer categories than the original data.

We overcome these difficulties by matching households in the regular EPA with the individual identifiers in the EF. Our problem is very different from the one described in Peracchi and Welch (1995), who match individuals across adjacent waves of the U.S. Current Population Survey. What we are doing instead is matching two different cross-sectional versions of the same data set.

**Table 7.** Tests for the equality of quarterly transitions probabilities of full-time respondents and other sample participants

Type of transition	Average significance level of the test			
	By period of entry		By province	
	Unweighted mean (s.e.)	Weighted mean (s.e.)	Unweighted mean (s.e.)	Weighted mean (s.e.)
FT to FT	0.2886 (0.300)	0.2821 (0.279)	0.0706 (0.135)	0.0387 (0.092)
FT to PT	0.3355 (0.308)	0.3352 (0.307)	0.1654 (0.259)	0.1948 (0.257)
FT to U	0.4211 (0.283)	0.4263 (0.282)	0.2304 (0.296)	0.1909 (0.255)
FT to O	0.3248 (0.278)	0.3334 (0.280)	0.2144 (0.298)	0.1239 (0.202)
PT to FT	0.0450 (0.099)	0.0408 (0.090)	0.1014 (0.197)	0.2468 (0.268)
PT to PT	0.0503 (0.114)	0.0416 (0.080)	0.0550 (0.154)	0.1525 (0.218)
PT to U	0.4926 (0.296)	0.5538 (0.297)	0.3067 (0.310)	0.3718 (0.283)
PT to O	0.2914 (0.298)	0.3580 (0.291)	0.3132 (0.295)	0.4613 (0.290)
U to FT	0.1326 (0.236)	0.1445 (0.251)	0.1843 (0.255)	0.3624 (0.327)
U to PT	0.3943 (0.312)	0.4276 (0.296)	0.3426 (0.300)	0.4624 (0.309)
U to U	0.1398 (0.206)	0.1682 (0.235)	0.0754 (0.165)	0.2887 (0.301)
U to O	0.3965 (0.317)	0.4578 (0.314)	0.1674 (0.263)	0.3077 (0.283)
O to FT	0.3308 (0.314)	0.3442 (0.306)	0.2497 (0.295)	0.3091 (0.295)
O to PT	0.5063 (0.306)	0.5073 (0.297)	0.3900 (0.310)	0.4809 (0.320)
O to U	0.2927 (0.303)	0.3180 (0.316)	0.2038 (0.296)	0.3642 (0.281)
O to O	0.2862 (0.302)	0.3056 (0.303)	0.1295 (0.215)	0.2809 (0.301)
All	0.2953 (0.305)	0.3151 (0.309)	0.1999 (0.276)	0.2897 (0.299)

The key variable for matching is the individual identifier. This is a 18-digit number, unique for each person, which consists of three parts: the census tract identifier, the household identifier, and the person identifier within the household. We employ a sequential procedure whose main steps are:

1. Sort the records in both sources using the common variables available. This matches the unique common records in the two files and helps identifying the census tract to which an individual (or household) belongs.
2. Sort the household identifiers within a census tract and the individual identifiers within each household, which fully matches both data sets.

In this way, we managed to successfully match 100% of the EF with the corresponding EPA records. We then added the individual identifiers from the EF to the regular EPA. So, henceforth we will only rely on our matched EPA files. The program that matches both data sets is available from the authors upon request.

## References

- van den Berg, G.J., Lindeboom, M. (1998) Attrition in Panel Survey Data and the Estimation of Multi-State Labor Market Models. *Journal of Human Resources* 33: 458–478
- Bover, O., Arellano, M., Bentolila, S. (1996) Unemployment Duration, Benefit Duration and the Business Cycle. *Economic Studies* 56: Bank of Spain
- Alba-Ramírez, A. (1997) Labor Force Participation and Transition of Older Workers in Spain. Working Paper 97–39. Universidad Carlos III, Madrid
- Alba-Ramírez, A. (1998) Re-employment Probabilities of Young Workers in Spain. *Investigaciones Económicas* 22: 201–224
- Beckett, S., Gould, W., Lillard, L., Welch, F. (1988), The Panel Study of Income Dynamics After Fourteen Years: An Evaluation. *Journal of Labor Economics* 6: 472–492
- Horowitz, J.L., Manski, C.F. (1998) Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations. *Journal of Econometrics* 84: 37–58
- Instituto Nacional de Estadística (1991) *Encuesta de Población Activa: Descripción de la Encuesta, Definiciones e Instrucciones para la Complementación del Cuestionario*, Madrid
- Instituto Nacional de Estadística (1994) *Encuesta de Población Activa: Tratamiento de la Información*, Madrid
- Instituto Nacional de Estadística (1995) *Evaluación de la Calidad de los Datos de la Encuesta de Población Activa: 1995*, Madrid
- Instituto Nacional de Estadística (1996) *Estadística de Flujos: Metodología*, Madrid
- Instituto Nacional de Estadística (1999) *Incidencia en los Trabajos de Campo en la Encuesta de Población Activa 1997*, Madrid
- Jiménez-Martín, S., Peracchi, F. (1999) La Calidad de la EPA en la Estimación de Transiciones en el Mercado de Trabajo. *Ekonomia* 43: 159–187
- Little, R.J.A., Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. Wiley, New York
- Paull, G. (1997) Dynamic Labor Market Behavior in the British Household Panel Survey: The Effect of Recall Bias And Panel Attrition. Oxford Labor Market Consequences of Technical and Structural Change, Discussion Paper series no 10
- Peracchi, F., Welch, F. (1995) How Representative are Matched Cross-Sections? Evidence from the Current Population Survey. *Journal of Econometrics* 68: 153–180
- Villagarcía, T. (1995) Existe un Sesgo de Inactividad en la Encuesta de Población Activa?. Documento de Trabajo 95–04, Univ. Carlos III de Madrid
- Zabel, J.E. (1998) An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior. *Journal of Human Resources* 33: 479–506



Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.