SOEP The German Socio-Economic Panel study

429

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP — The German Socio-Economic Panel study at DIW Berlin 2017

The measurement of labor market entries with SOEP data. Introduction to the variable EINSTIEG_ARTK

Marco Giesselmann, Paul Schmelzer, Mila Staneva, Tobias Wolfram

DIW SOEP

Running since 1984, the German Socio-Economic Panel study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing. The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at **http://www.diw.de/soepsurveypapers**

Please cite this paper as follows:

Marco Giesselmann, Paul Schmelzer, Mila Staneva, Tobias Wolfram. 2017. The measurement of labor market entries with SOEP data. Introduction to the variable EINSTIEG_ARTK. SOEP Survey Papers 429: Series D. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)
DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany
soeppapers@diw.de

The measurement of labor market entries with SOEP data. Introduction to the variable EINSTIEG_ARTK

Marco Giesselmann
Paul Schmelzer
Mila Staneva
Tobias Wolfram

Summary:
The variable EINSTIEG_ARTK in the file BIOJOB indicates the wave of a first labor market entry. The variable refers to the entry into the labor market after the completion of education and apprenticeship. Since we account for all kinds of employment (full-time, part-time and marginal employment), such a definition captures the uncertainties and instabilities associated with the early career phase. At the same time, side-jobs or apprenticeships are explicitly assigned to the educational phase and excluded from the concept of labor market entry. Among the several plausible concepts and operationalization of labor market entries with SOEP data (among them the “age of first job”-variable), we consequently hold this indicator particularly suited for scholars who want to study the impact of labor market institutions on early career outcomes. As an alternative for researchers who do not rely on exact information on the transitional wave, we offer an alternative variable EINSTIEG_PBIO (that covers more cases, but is only exact on the basis of years, not waves).

Measuring labor market entry with the SOEP

Entry to the labor market is a key transition in a person's life course in modern societies. Therefore, life course researchers aim at defining and measuring labor market entries in an accurate way. However, not just life-course researchers are focused on labor market entries: In many standard applications of empirical social science research entrants are defined as a specific social group, whose economic and subjective living conditions shall be compared with those of other labor market groups. For such studies, similarly, using an appropriate operationalization of labor market entries is essential.

The SOEP offers exhaustive and detailed information on labor market status and biographies. However operationalizing the transition from the educational system to the labor market is a complicated issue. There are two reasons for that: For one, a person's labor market entry has, unlike other socio-demographic events like divorce, childbirth, or unemployment, often a gradual nature: where to set the transitional year or wave in such cases? Furthermore, knowing *the year* of a transition doesn't mean knowing *the wave* in which the transitions occurred (as the transition might have taken place before or after the interview in the transitional year). The same, of course, is also true vice versa. Thus, a researcher aiming at defining the measurement of transition (*t0*) in a life-course or life-event design needs a different, more refined operationalization than a researcher who seeks to model the duration of certain life-phases on the basis of retrospective data.

Accordingly, there are several ways to define and to measure labor market entries. In view of the gradual, sometimes even recursive character of the transition from education to employment, the determination of an entry point is strongly linked to theoretical concerns. Shall phases of vocational training be considered as last part of the educational phase or first part of the employment biography? Shall jobs that end with a re-entry into training be considered as labor market entry or an interruption of the educational phase? There is no perfect answer to such questions. Rather, the definition of "labor market entry" has to be aligned with the concepts and concerns underlying the research question. However, we suggest a concept and an operationalization of an entry-variable which might be useful in many empirical analyses involving labor market entries. This suggestion is accompanied by the provision of two easy implementable, generated variables which are now added to the BIOJOB file.

One variable, EINSTIEG_ARTK, is constructed on the basis of prospective data and monthly calendar data. It hints at the wave, in which the transition has been recorded. Therefore, it is designed for life-event and life-course analysis, in which the identification of the transitional wave (t0) is essential.

The other variable, EINSTIEG_PBIO, is based on prospective and retrospective data and hints at the year, in which a labor market transition took place. It is not as exact as the EINSTIEG_ARTK, does not necessarily point t0, but also covers transitions outside the

prospective observational window of a person; in cases, in which prospective information related to the labor market entry is not needed, it provides larger samples than EINSTIEG_ARTK.

In what follows, we describe the operationalization and background of these two variables in detail. We also present some descriptive statistics on the basic properties and comparability of these variables. In this comparison, we also include the self-reported variable AGEFJOB (*age of first job*).

EINSTIEG_ARTK: Identifying the entry-wave on the basis of prospective and calendar data

This generated variable EINSTIEG_ARTK[1] indicates the year of the annual measurement (or, rather, *the wave*), in which the first entry in full-time, part-time or marginal employment after the completion of education and apprenticeship is detected. It refers only to labor market entries within the prospectively observational window of a person and is generated mainly using biography data from the calendar dataset ARTKALEN. This data provides monthly information on activity status since an individual entered the SOEP and is surveyed within the yearly interviews. In cases where the calendar data does not reveal the first labor market entry, annual information from the personal questionnaire (a variable indicating the self-reported entry in employment stemming from the annual datasets $PGEN) is used in order to identify possible "hidden" entries within the observational window. To validate the information on first entry, additionally a variable indicating the age at first job stemming from the biography interview is used.

Compared to existing measures of labor market entry in the SOEP, the most important of which is "age at first job" (AGEFJOB)[2], the variable EINSTIEG_ARTK offers some enhancements:

a) It is based on a more clear and consistent definition of what a first labor market entry is. Here the first labor market entry is conceptualized as the entry in the first job after the completion of (secondary and tertiary) education and apprenticeship. The variable AGEFJOB, on the other hand, captures labor market entries at very different stages of the educational and employment biography. One reason for this is that it largely relies on a self-assessment of what a labor market entry is. In the biography interview all respondents are asked when they first started to work and this leads to very diverse self-reported labor market entries ranging from the first side-job in high school to the first full-time stable employment matching the own professional field.
b) Another reason is that by generating the variable AGEFJOB for one large part of the respondents a different operationalization strategy is applied – for people who have never been employed at the time point of the Biography Interview the very first

[1] An equivalent operationalization of labour market entry has been used in Giesselmann (2015): Differences in the Patterns of in-work Poverty in Germany and the UK. European Societies 17(1): 27-46.
[2] A detailed description of the generation process of AGEFJOB can be found in the biography documentation of the SOEP, which is updated on a yearly base.

observed labor market entry from the spell-data PBIOSPE is used for the generation of the "age at first job"-variable. This also leads to inconsistent labor market entries since this generating strategy often captures student side-jobs. In order to account for such shortcomings we rely only on observed entries and use self-reported information only for verification and plausibility checks.

c) Additionally, the new EINSTIEG_ARTK variable refers to the entries that precede the year of interview. While some entries occur in the same year before an interview takes place (e.g. the job starts in February 2000 and the interview takes place in June 2000), other entries occur at the same year after the interview took place (e.g. interviews takes place in June 1999 and the job starts in September 1999). In the latter case we shift the job entry variable to the following year (in our example we assign the entry year 2000 for a person who started their first job in September 1999). Such a strategy allows the utilization of information (e.g the personal questionnaire) at the time of interview that follows the job entry.

The following paragraphs describe the operationalization of the concept in detail. As mentioned before, the activity calendar (ARTKALEN) is used to observe the transition from education to employment. The labor market entry is defined as the first month in full-time, part-time or marginal employment after the end of the educational phase of a person. The educational phase generally ends with the last educational spell of a person. However, only such educational spells qualify as a potential last spell, which are not preceded by a phase of non-education of more than 24 months. This means, if a person studies for four years, leaves the educational system for more than 24 months and then re-enters education, the end of the first four-year educational spell will be taken as the end of the educational career. Important is that people in vocational training are also considered to be in education. For such people the variable indicates the transition from vocational training to the labor market as labor market entry. "Further training/ retraining" spells ("Weiterbildung"), however, are, by contrast to "vocational training" spells ("Ausbildung") not considered as part of the educational phase. Actually, this last rule applies only from 1999 on, because before this time "Further training/ retraining" and "Vocational education" were summarized in one single spell type. This means that we were not able to disentangle the "Further training"-spells from educational phases prior to 1999. Therefore, being in "Further training/ retraining" accounts for participation in the educational system before the year 1999, but not afterwards. This might be an important restriction in the consistency of the underlying conceptualization and operationalization of the entry-variable: for those users, who perform analyses of entrants with a strong focus on sensitive trends, we therefore suggest to use only data from 1999 onwards.

In some cases, an actual labor market entry lies before the observational window (first year of interview), although the pattern within the observational window indicates (falsely) a labor market entry. In order to account for such errors, the identified entry is compared with the self-reported first entry in employment from the biography questionnaire. In this way we identify respondents who report a first labor market entry in the biography questionnaire

that occurred before they entered the SOEP. If there is such ambivalent information, we have to doubt whether the observed transition from education to employment in the activity calendar (ARTALEN) actually captures the first entry into the labor market – it may wrongly capture re-entries in the labor market after extended education or on-the-job training. For this reason we do not consider observed entries in employment in ARTKALEN, when the respondents report a first entry in employment taking place before the observational window and a) when they were not in education at this point of time or b) were in education, but were older than 27. To check whether the people were in education at the time point of the self-reported labor market entry we used information from the dataset PBIOSPE which contains the self-reported biography since the age of 15.

In some other instances, the observational window of a person captures the labor market entry, but not the preceding educational phase. This happens, for example, if a person finished education several years before the first interview, then is unemployed, starts being surveyed during unemployment and then starts a job, which, according to our conceptualization, would count as labor market entry. The above described generating strategy with ARTKALEN, however, does not identify such entries. For this reason, we use annual information on labor market entry from the variable “occupational change” (from the files $PGEN, indicating whether a person is first time employed) when the monthly biography in ARTKALEN does not show a transition from education to employment. In such cases, we assign a labor market entry when a person reports first time employment in the yearly interview is actually employed at this time and a) is not in education and younger than 27, b) is enrolled in a university and older than 27, c) is attaining vocational training and older than 25. If these conditions are not met, the entry indicated in the variable “occupational change” is shifted to the first following year, in which they are met. Similar to the ARTKALEN generation strategy, the entries based on the variable “occupational change” are compared with the self-reported entries in the labor market from the biography interview. When a self-reported entry occurs before the observational window and the person was not in education at this time, the labor market entry based on the variable “occupational change” is considered as not valid.

Another relevant factor we take into account is conscription, which was in place in Germany until 2011, with a length ranging from 6 months to 20 months in different time periods and places (East and West Germany). Therefore in some cases we observe respondents finishing their secondary education, having a gap of some months, then being conscripted and after again some months of gap time beginning their tertiary education. In such cases the overall time between the ends of the two education periods can easily span 24 months or more, so our algorithm would misidentify the last phase of education and hence the first job as theoretically desired. To avoid this we also check for spells which denote conscription and control whether there are further educational spells at most one year after finishing conscription.

Using the outlined methodology we are able to reconstruct the job market entry-years of roughly 7500 respondents on the basis of SOEP v32, released in 2016.

EINSTIEG_PBIO: Reconstructing Job Entries using PBIOSPE

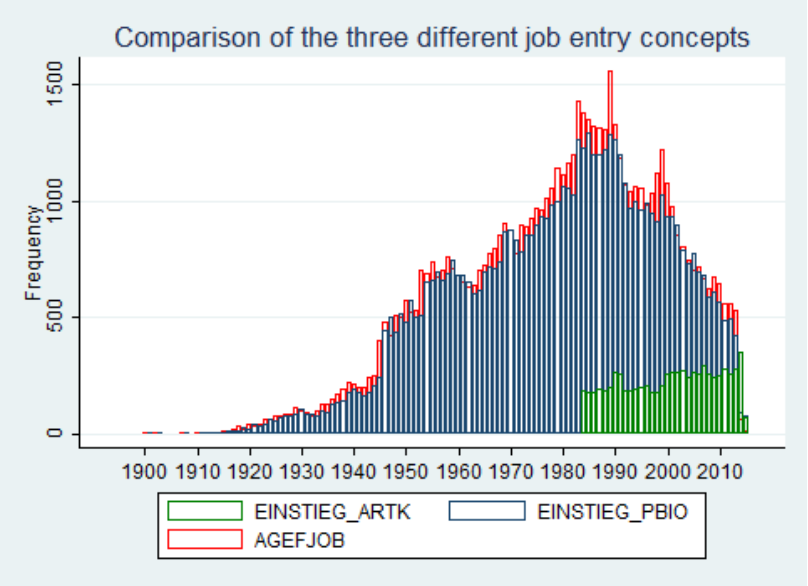
As mentioned in previous section **EINSTIEG_ARTK** entails about 7500 of respondents entering their first job (on the basis of SOEP v32). This low number is explained by the fact that to be identified by our algorithm the beginning of the job (and the end of the educational) biography of a SOEP-participant has to be part of the ARTKALEN-dataset. This is only the case if one became part of the survey in late youth or early adolescence and did not leave the sample before a first employment could be observed, so only for a fraction of first jobs as defined by EINSTIEG_ARTK actual transitions can be identified. Naturally, the earliest entries therefore begin in 1984. But as the problems concerning the validity of the AGEFJOB-variable, which we elaborated at the beginning, hold true for all observations: Currently more than 60000 valid responses to this variable are listed in BIOJOB, the earliest indicating a job entry in the year of 1900 (!). To offer a compromise between the likely inconsistent AGEFJOB-data and the likely correct EINSTIEG_ARTK-observations we created a second variable. Instead of using information from ARTKALEN, EINSTIEG_PBIO employs the dataset PBIOSPE, which includes retrospective spell-data gathered from the Biography Questionnaire.[3] For understandable practical reasons though this data is just available on a yearly and not a monthly basis like it is the case with ARTKALEN. The implied loss of granularity induces a potential higher risk of misclassifications compared to EINSTIEG_ARTK – specifically if rather *the wave* of entry instead of *the year* is the focus of interest – while enabling us to reconstruct job entries for a vastly higher amount of respondents, namely almost everyone who ever filled out the biography questionnaire – which is true for the majority of the respondents ever surveyed in the SOEP. Still, we should keep in mind that the benefits from the substantially larger number of job entries compared EINSTIEG_ARTK may not be overestimated: For most identified first jobs which fall in the time frame before the person became part of the sample and therefore have a record in EINSTIEG_PBIO but not in EINSTIEG_ARTK, there is no prospective information on any covariate in the year of entry. Therefore, from many types of analyses, the overlapping entries will be eliminated due to lack of information, anyway.

The basics of the generation of EINSTIEG_PBIO are similar to EINSTIEG_ARTK with minor restrictions: Firstly, as the categorization of activity spells in ARTKALEN is more granular than that in PBIOSPE certain checks which apply to exclude further education/retraining-spells and make it possible to divide them from other episodes of education (as mentioned above) are not available, but comparisons indicate that the differences induced by this shortcoming are miniscule. Secondly, as just explained, it is not the main purpose of EINSTIEG_PBIO to merge the identified entry-year with covariates from the affiliated survey year, which most of the time just do not exist anyway. Therefore the given year is really just the year of the identified entry and does not refer to the respective wave, the earliest yearly measurement after the transition.

[3] Note while retrospective information is used to reconstruct the spells of PBIOSPE before participants take part in the survey ARTKALEN is used to construct the spells of PBIOSPE for the following years.

Comparisons of EINSTIEG_ARTK, EINSTIEG_PBIO and AGEFJOB

To gain a better understanding of the pitfalls of the three different variables we will show example cases for which their estimated job entry years differ. For an accurate comparison we generated job entries based on information of year of birth and AGEFJOB. We start by a descriptive comparison of the frequency plots of the different variables:



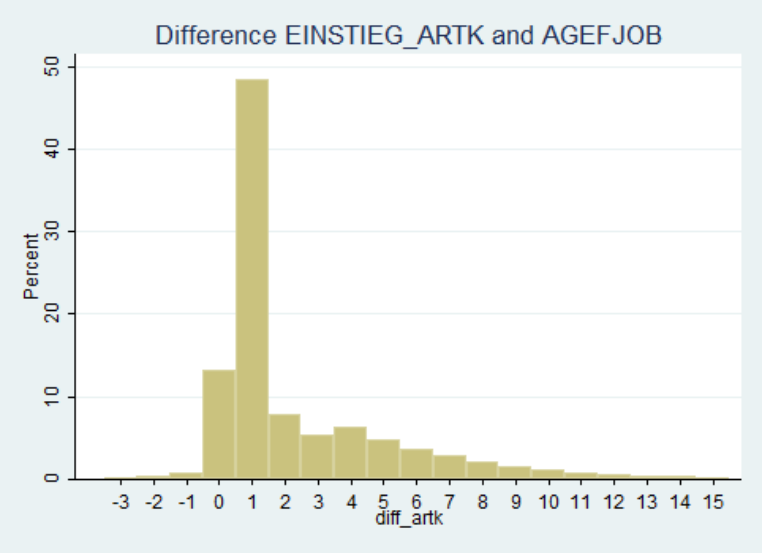
As the graph above demonstrates we are able to catch most of the AGEFJOB-cases using EINSTIEG_PBIO, while only a minority of the cases can be reconstructed using EINSTIEG_ARTK. Still all three variables are highly correlated as one would expect:

	AGEFJOB	**EINSTIEG_ARTK**	**EINSTIEG_PBIO**
AGEFJOB	1		
EINSTIEG_ARTK	0,9532	1	
EINSTIEG_PBIO	0,9912	0,9848	1

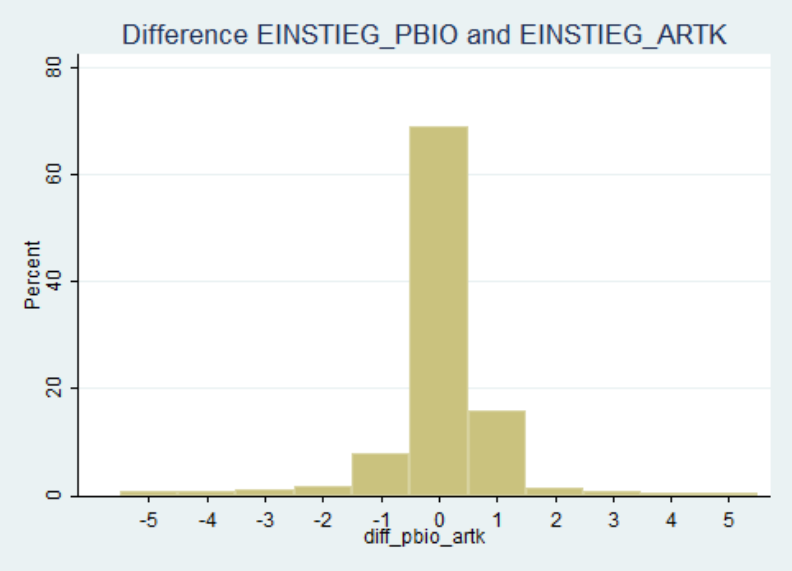
However, even slight differences in the assignment of an entry measure can be substantial, as can be seen if we plot the differences between EINSTIEG_ARTK and AGEFJOB:[4] for most cases we observe an antedating of one year, which is due to the fact explained above: EINSTIEG_ARTK estimates not the direct year of job entry but points to the year of the earliest yearly measurement after the transition. In the same vain this also explains differences between EINSTIEG_PBIO and EINSTIEG_ARTK: If a labour market entry in a given year occurs after the interview, then there is a slight difference between the variables.

[4] This is done by generating the year of first job using AGEFJOB as the birth year of each observation and then subtracting this value from EINSTIEG_ARTK. A positive difference therefore means that EINSTIEG_ARTK indicates a later job entry than AGEFJOB.

This supports the idea that a significant share of people systematically interpret the "age at first job"-question in a different way than the majority.



At the same time the differences between EINSTIEG_ARTK and EINSTIEG_PBIO are minor:[5]



As mentioned above we now want to explore archetypical individuals to explain the reasons for these differences. We start by looking at respondent *82403*, who has a rather complex job-biography, looking at the ARTKALEN-data:

[5] We subtract EINSTIEG_PBIO from EINSTIEG_ARTK, therefore a positive value indicates that the value of the latter is higher than that of the former.

Spelltype	Begin	End
School, College	1993 Jan	1995 Dez
Housewife, Husband	1996 Jan	1996 Dez
Vocational Training	1996 Jan	1996 Dez
Unemployed	1997 Jan	1997 Jul
Vocational Training	1997 Aug	1998 Sep
Unemployed	1998 Okt	1999 Aug
Full-Time Employed	1999 Sep	2002 Jul
Housewife, Husband	2000 Jan	2001 Dez
Unemployed	2002 Jul	2002 Okt
Part-Time Employment/Marginaly Employed	2002 Nov	2004 Feb

This person finished his school in December 1995 and started his vocational training in the beginning of the next year while still denoting himself as husband. After that he is unemployed for half a year after which he again starts or continues a vocational training for almost a year and which ends in September 1998. This spell is followed by another period of unemployment. After that he finally starts working in September 1999 which by any means would count as the first job and is in agreement with AGEFJOB. PBIOSPE gives a similar though less detailed impression:

Spelltype	Begin	End
School, College	1992	1995
Vocational Training	1994	1994
Housewife, Husband	1996	1996
Vocational Training	1996	1998
Unemployed	1997	1999
Full-Time Employment	1999	2002
Housewife, Husband	2000	2001
Unemployed	2002	2002
Part-Time Employment	2002	2006
Housewife, Husband	2003	2003

And while AGEFJOB and EINSTIEG_PBIOSPE identify 1999 as the year of job entry, EINSTIEG_ARTK gives 2000, as this year where the relevant survey data was gathered. Therefore everything works as expected in this case.

Contrast that to person *107303*, whose self-reported first job (AGEFJOB) is set in 1982 while both versions of EINSTIEG predict 1992, as it is elaborated below. A 10 year difference sounds unrealistic, but looking at the data we see that EINSTIEG is working as we would expect it:

ARTKALEN:

Spelltype	Begin	End
School, College	1983 Jan	1991 Aug
Part-Time Employment / Marginally Employed	1983 Jan	1991 Aug
Housewife, Husband	1984 Jul	1984 Dez
Housewife, Husband	1987 Jan	1987 Dez
Full-Time Employmet	1991 Sep	1993 Jul
Housewife, Husband	1993 Jan	1993 Dez
Maternity Leave	1993 Aug	1993 Dez

PBIOSPE:

Spelltype	Begin	End
School, College	1978	1991
Part-Time Employment	1982	1991
Housewife, Husband	1984	1984
Housewife, Husband	1987	1987
Full-Time Employment	1991	1993
Housewife, Husband	1993	1993
Other	1993	1993

During her studies *107303* started to work part time in 1982 as we can see in PBIOSPE, but did not start to work in a real job before 1991. That is also supported by looking in the detailed job information: While in 1990 she is listed as office helper in the StaBua-variable, which denotes job by the classification system of the federal statistical office, the next year her profession changed to sports teacher. EINSTIEG_ARTK therefore gives 1992 as the year of the relevant survey data and EINSTIEG_PBIO gives 1992 as this is the first year of her job after she finished her studies.

A good example for difficulties and ambiguities which arise and are impossible to fix is delivered by observation *569505*:

ARTKALEN:

Spelltyp	Begin	End
School, College	1992 Jan	1994 Feb
Unemployed	1994 Mar	1994 Dez
Full-Time Employment	1995 Jan	1995 Jul
Unemployed	1995 Aug	1996 Mar
School, College	1996 Jan	1996 Mai
Full-Time Employment	1996 Apr	1997 Feb
Unemployed	1997 Mar	1998 Jan

PBIOSPE:

Spelltype	Begin	End
School, College	1992	1994
Unemployed	1994	1998
Full-Time Employment	1995	2003
School, College	1996	1996

By his own account (AGEFJOB) he started working in 1995, but there was an educational phase after that (in 1996) and therefore the year after the education is taken as the first year of work for EINSTIEG_PBIO (though the education period was rather short) and the year of survey is taken for EINSTIEG_ARTK, therefore the dates coincide. But which date is the correct entry? Determining the year of job entry in such cases is a bit arbitrary and never perfect.

Sometimes AGEFJOB exists, while both EINSTIEG variables are missing: This is explained by the example of case *285104*, as he reported the sidejob during his studies as his first job (therefore AGEFJOB gives 1988 as year of first job) but left the survey before ever entering a real job:

ARTKALEN:

Spelltype	Begin	End
School, College	1984 Jan	1987 Jul
Military, Civil Service	1987 Aug	1988 Sep
School, College	1988 Okt	1989 Jul
Part-Time Employment / Marginally Employed	1988 Dez	1988 Dez

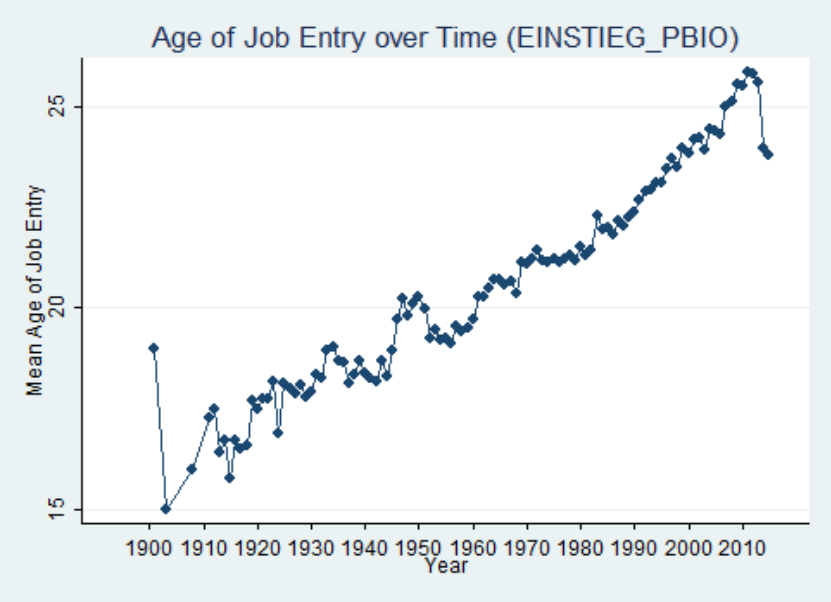
PBIOSPE:

Spelltype	Begin	End
School, College	1984	1994
Military, Civil Service	1987	1988
Part-Time Employment	1988	1989
Part-Time Employment	1991	1994

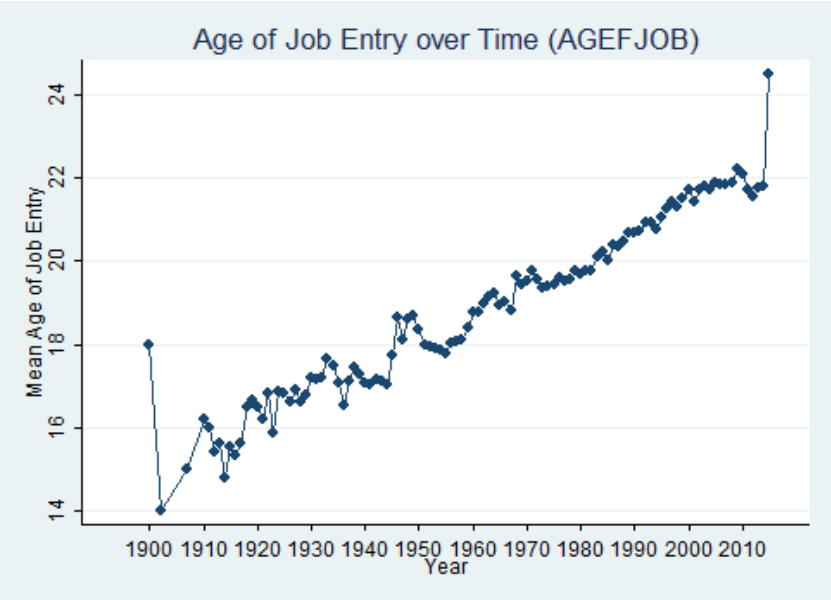
Therefore there is no chance to ever verify his real job entry.

Descriptive Statistics on Labor Market Entrants

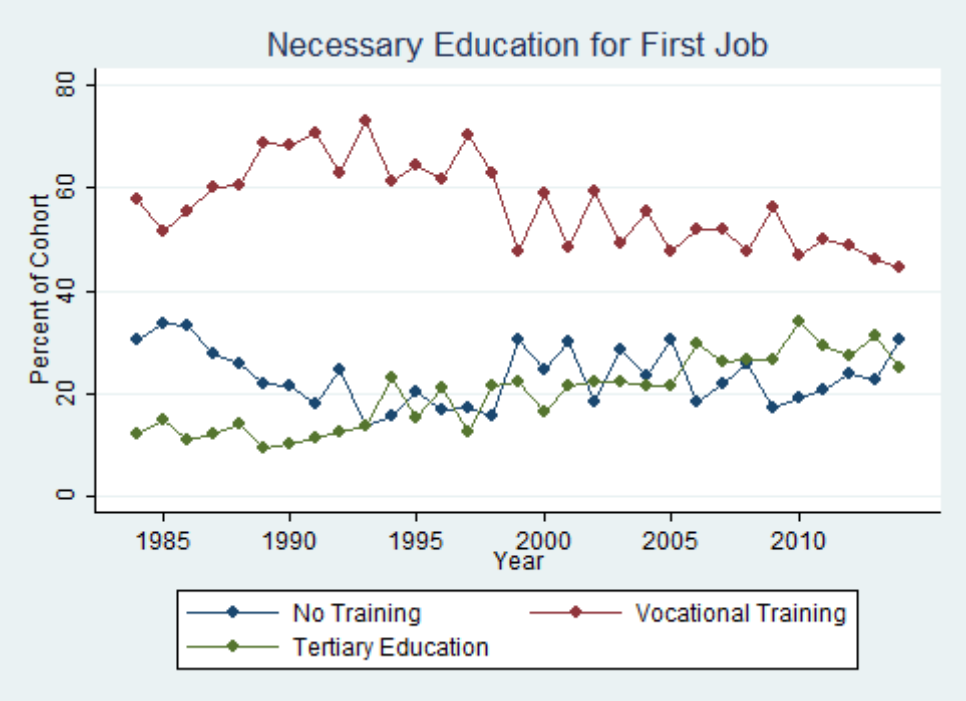
So far we compared the different variables and their concepts. In the last part of this documentation we want to give a short overview over some descriptive information on labor market entrants. An interesting question in that regard is the mean age of job entry over time. Considering the tremendous amount of economic and social change which happened during the last century it is likely that the age of job entry over time should have risen.



As we should not take the outliers in the first and last years to seriously as there is just a low amount of observations available for these time periods we see a clear linear pattern in support of our assumption. If we compare this to the same graph based on AGEFJOB-data we see the consistent biasedness of the self-reported year of first job entry. Especially during the last thirty years AGEFJOB underestimates the mean age of job entry.



Another interesting information is the required level of education for the first job as asked in $p and its fluctuation over time. As there are in fact eight different categories[6] we group them together to end up with just three to guarantee readability of the plot[7].



"Vocational Training" and its grouped surrogates dominate the picture, while "Tertiary Education" and "No Training" are close to each other since the early nineties. Still, there seems to be no definitive trends.

[6] No training, vocational training, college degree, university degree, short briefing at workplace, longer vocational adjustment, attendance at courses/seminars, institution of higher education

[7] No training and short briefing at workplace are grouped as "No Training", vocational training, longer vocational adjustment and attendance at courses/seminars as "Vocational Training" and the rest as "Tertiary Education"