Deriving Educational Attainment by combining data from Administrative Sources and Sample Surveys

Recent developments towards the 2011 Census

Frank Linder & Dominique van Roon

Abstract: Data on educational attainment are crucially important in socio-economic research and government policy. Multivariate analyses often show that the impact of education in explaining a variety of social phenomena is undeniable. For example, well-educated people tend to have better prospects on the labour-market. This knowledge may convince governments of the need to prevent students from dropping out of school and to invest more in education programmes. Also, several studies indicate that higher levels of education tend to generate more average income. Altogether it is not surprising that educational attainment is a standard variable in the Census Programme.

The Virtual Census of 2001 in the Netherlands used the Social Statistical Database (SSD) as its key source. The SSD can be considered as a statistical framework that provides information on demographic and socio-economic issues. It is constructed by micro-linkage and micro-integration of several administrative registers and household sample surveys, which ensures coherence, consistency and completeness of the SSD data. The SSD gives priority to administrative data sources whenever they are available. Sample surveys are explored to compensate for information that is not (yet) in registers. At the time of the 2001 Census the only source providing data on educational attainment was the sample of the Labour Force Survey (LFS).

In the last decade a wide variety of administrative education registers has come at the disposal of Statistics Netherlands. This has increased the information content in the field of education considerably. So, it is now possible to produce more reliable estimates on education level than the LFS results from earlier years, in particular when smaller populations are involved. It is a development of which the next Census of 2011, with its detailed table programme, will definitely take advantage.

As the education registers do not cover the entire Dutch population, the LFS still plays an important part in filling the gaps. Most older citizens completed their education career before registers came into consideration for official statistics. Besides, the education registers have no information on studies abroad. In these cases the LFS compensates for the lack of information in the administrative sources.

Deriving educational attainment by combining data from registers and the LFS supplement is far from simple. Problems such as the complex sampling design for the combined register and sample data have been solved by applying sophisticated methodology. At present we consider revising the weighting strategy. In addition bootstrap-methods are being developed for the construction of variance estimators and confidence intervals, to determine if the estimation of education levels is accurate enough for dissemination. The results of this research will be incorporated in the production process before the actual data preparation for the Dutch Virtual Census 2011 is due to start.

Keywords: accuracy; administrative register; bootstrap-method; calibration; consistency; educational attainment; micro-integration; micro-linkage; the Netherlands; population census; repeated weighting; sample survey; sampling design; scaling; Social Statistical Database; weighting strategy.

1. Introduction

When Queen Beatrix opened the Dutch Parliament in 2007[1], she said: „Education is the steppingstone to success in life“. A year later she added to this: „Training and education increase one’s opportunities. The government works hard to improve education .....Tackling the problem of school drop-out remains a priority“ (Queen’s speech, 2007 and 2008).

[1] On the third Tuesday of September (Prince’s day), the Queen of the Netherlands opens parliament with a traditional speech in which she outlines government policy.

The authors thank Bart Bakker, Paul Knottnerus, Léander Kuijvenhoven, Sander Scholtus and Eric Schulte Nordholt for their valuable comments on an earlier draft of this paper.

Similar statements were made in the EU Lisbon Strategy[2]: „Education and training policies are central to the creation and transmission of knowledge and are a determining factor in each society's potential for innovation.....In addition, the positive impact of education on employment, health, social inclusion and active citizenship have already been extensively shown" (CEC, 2003).
It is clear that education is high on the political agenda. And with reason, because in socio-economic research educational attainment is often confirmed as a key social indicator explaining a wide range of social phenomena. Education can be considered as an investment in 'human capital', which may generate economic benefits, such as a higher labour force participation and higher earnings. See e.g. Becker (1993) and Mincer (1974) on human capital theory.

To set out policy and to conduct research, good quality data are indispensable. In the Netherlands statistical information on educational attainment was exclusively the domain of the Labour Force Survey (LFS) for a long time, except for the ten-year population censuses in the past. It was not until the new millennium that administrative registers on education became available for statistical purposes. In that respect the Netherlands lags behind the Nordic[3] countries, which have a long experience with register-based statistics. For education, as far back as the seventies and eighties (UNECE, 2007).

Until the seventies the Dutch population census was organized in the traditional way, that is by field enumeration. Administrative sources, the Population Register to begin with, became more and more important for statistics. The first census in the new century, the Virtual Dutch Census of 2001 relied heavily on administrative data. The key source for this census was the Social Statistical Database (SSD), a statistical framework of micro-linked and micro-integrated administrative registers and sample surveys, see Linder (2004) and Schulte Nordholt & Linder (2007).

The 2001 Census statistical information on education had to be drawn from the sample of the LFS. This is going to change with the Census of 2011. The recent introduction of administrative sources on education allows the construction of a census variable educational attainment, which is based on registers for a substantial part. However, for the remainder the variable will still depend on the LFS, particularly for those people who completed their education prior to the administrations. Nevertheless, this development is an important step forward because more reliable estimates on education levels will be produced than ever before as result of the increased number of records. This will certainly be the case in detailed tables of the Census Programme (UN, 2008).
The innovating aspect in this approach of micro-integration is that data are combined from registers as well as from sample surveys for one and the same variable. In the past, data sources for a variable in an integration framework like the SSD were either registers or sample surveys, but never both at the same time. Combining administrative and sample data for the same variable gives the statistician some thorny methodological issues to solve. These problems include the weighting mechanism and validity of the data in case the information is out-of-date.

The next section presents the Social Statistical Database in the past and the present and its role in the Census. Section 3 describes the sources that are explored for the variable educational attainment. Section 4 deals with the micro-integration aspects of deriving educational attainment. Section 5 goes into the weighting strategy. Section 6 focuses on the accuracy of the results. Section 7 completes the paper with some concluding remarks.

2. The Social Statistical Database in the past and the present

For the 2001 Census Programme, Statistics Netherlands made use of a significant development at the time, the creation of the Social Statistical Database (SSD). The SSD is the final product of micro-linkage and micro-integration of source files containing demographic and socio-economic data. The advantage of the SSD over the constituent sources is that it is a coherent and consistent integration framework, which gives a more complete coverage of the target population. These qualities do not apply for provisional SSD figures on current information, see below.

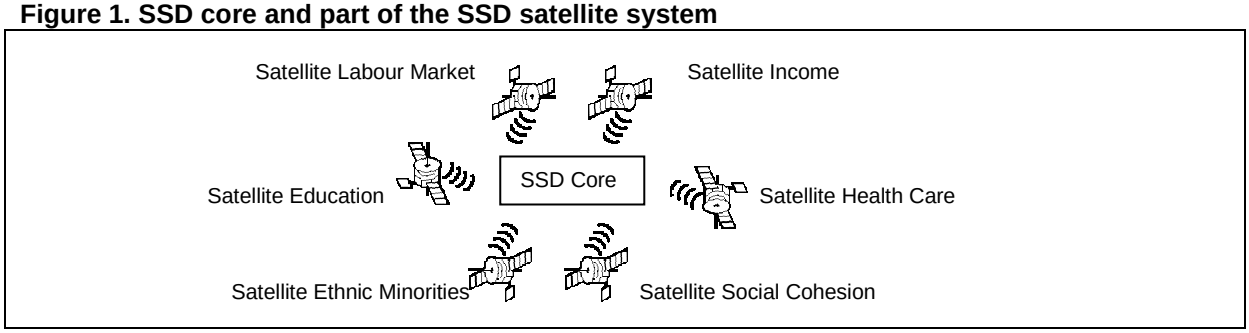
[2] At the Lisbon European Council held in March 2000, the Heads of State and Heads of Government launched the so-called Lisbon Strategy, aiming at making the European Union (EU) „the most dynamic and competitive knowledge-based economy in the world".
[3] Denmark, Finland, Iceland, Norway and Sweden.

Primarily the statistical information in the SSD comes from administrative registers. These registers are a relatively new phenomenon of the last decade or so. Although they originally served administrative purposes, their quality often turns out to be sufficient for statistical purposes as well[4]. Statistics Netherlands has substituted more and more sample surveys by administrative registers. The Survey on Employment and Earnings, for example, has vanished completely and was replaced by a register with information on jobs and wages.
Sample surveys, however, are still indispensable when information is not available in registers. A good example of this is the variable occupation, as there is no occupation register in the Netherlands.

Most of the registers are provided with a personal identifier for each citizen, the Citizen Service (CS) number[5]. The CS-number can be used as a linkage-key, however, for protection reasons it will be encrypted into a unique Record Identification Number (RIN-person). The Population Register (PR) with demographic information on every inhabitant of the Netherlands plays a central role. All the other data sources of the SSD are linked to this register, in other words the PR serves as the backbone of the SSD. After micro-linkage has taken place this set of files undergoes a micro-integration process with checks, harmonization, adjustment for incorrect data, and completion in case of missing data, see e.g. Van der Laan (2000).

At the time of the 2001 Census statistical information in the SSD was mainly confined to data about the population (e.g. sex, age, marital status), labour (e.g. jobs, wages, economic activity, place of work) and social security (e.g. benefits). However, the SSD continued in its development, and expanded its scope to include current themes such as enterprises, income, social cohesion, social dynamics, housing, regional information, social and spatial mobility, ethnic minorities, education, health care, and security.
The production of provisional figures for up-to-date SSD information is of recent date, but only for key variables. An important drawback of the SSD is that it cannot satisfy the need for current information because micro-integration is a time-consuming process. To speed up the process the provisional SSD figures are exclusively based on data in the main source. In other words, for up-to-date info we drop an important feature of the SSD, that is the requirement of consistency with other sources. However, the concept of the provisional variables is basically the same as that of their integrated counterpart.

Figure 1. SSD core and part of the SSD satellite system



With so many fields to cover, the SSD-system threatens to become unmanageable. Therefore, it is split into smaller organizational units, the SSD core and SSD satellites. The SSD core is the pivot in the system, and concentrates on demographic and socio-economic information which is relevant in almost any field, such as sex, age and socio-economic category. The SSD satellites each cover a specific theme, e.g. health care or social cohesion. The satellites and the core are consistent in information over all units. So, the objective of one figure on one phenomenon is guaranteed within the whole system of core and satellites.
The educational attainment variable is positioned in the SSD core because of its determining role in many social processes. This means that every SSD satellite implicitly has consistent information on the level of education.

[4] Bakker et al. (2008) and Daas et al. (2009) give an extensive treatment of the quality aspects of register data.
[5] Previously this was called the social security and fiscal (SoFi-) number.

The extension of the SSD with a wide variety of themes increases the possibilities for the next Census. When the Virtual Census of 2001 was produced, household information for example was not yet an integral part of the SSD. It had to be obtained from separate demographic files. Nowadays it is a fixed integrated item in the SSD core.
Something similar applies to such Census variables as attending educational institutions and educational attainment. In the 2001 Census time both variables were LFS variables, while today they are fully integrated within the SSD core. Moreover, the first variable is completely register-based and the second is so at least for a substantial part.

3. Sources

Educational attainment in the Netherlands has long been the exclusive domain of the Labour Force Survey. Nowadays, administrative education registers are a serious competitor for the LFS where education levels are concerned. Education registers offer additional value, in particular when detailed estimates are needed on smaller subpopulations. Besides, the registers are not troubled by the persistent problem of selective non-response that plagues the LFS. On the other hand, the LFS complies better with the guidelines for the 2011 Census, in the sense that 'all education which is relevant to the completion of a level shall be taken into account even if this was provided outside schools and universities' (CEC, 2009).

3.1. The Labour Force Survey

The Labour Force Survey (LFS) is an annual household sample survey that started in 1996 and includes well over 100 thousand individuals[6] living in private households, about 1% of the total population. The sample size can be considered sufficient for reliable estimates on educational attainment in larger subpopulations, such as the unemployed in the eastern region of the country or male foreigners in Rotterdam.
Problems may arise with populations on a smaller scale, small municipalities for example. Normally, this is solved by presenting outcomes based on the unified sample survey of two or three consecutive years. Standard errors will generally decrease as result of the higher number of observations. As long as the variable concerned is only changing slowly over time this is an acceptable method. Education levels are rather stable within a period of one or two years, and so are suitable for such an approach. For this reason the census variable educational attainment in the 2001 Census (reference date 1 January) was based on the LFS of 2000 and 2001.

The LFS provides the complete education career of a respondent until the date of the interview. Education programmes are coded in 6 digits according to the Standard Classification of Education (SCED2006), see Schaart et al. (2008). The Census Programme of 2011[7] (UNECE, 2006) recommends the use of the International Standard Classification of Education 1997 (ISCED1997), issued by UNESCO. That is no problem, because with their high degree of detail the SCED codes can easily be converted into ISCED levels.

Unlike most of the registers, the Labour Force Survey lacks a personal identifier such as the Citizen Service Number. Linkage of the LFS with other sources often occurs with a key that is built up by combination of the identifiers sex, date of birth, postal code and house number. When linked to the Population Register linkage rates still as high as 97 and 98 percent are achieved.

3.2. Administrative education registers

3.2.1. Primary education

There is no administrative register yet on primary education. The first Primary Education Number Register (school year 2008-2009) is expected in 2010. This will have no effect for the 2011 Census, as

[6] When the LFS is redesigned in the near future, the sample size may be reduced by some 15 percent.
[7] If a new ISCED classification is in force on 1 January 2011 the Guidelines of the Census Programme 2011 suggest to use it instead of ISCED1997.

the latest draft census guidelines (CEC, 2009) require educational attainment of people under the age of 15 years to be classified as 'not applicable'.

3.2.2. Secondary education

There are several administrative registers on secondary education.
For 1999 onwards there is a so-called Exam Results Register (ERR) with data of students taking exams in secondary general and in lower secondary vocational education. In addition to the exam results there is also information on the type and level of the education programme. The ERR codes of level of education can be transformed into SCED and ISCED codes.
An alternative way of getting information on secondary education graduates is to use registers of Higher Education (see below). These registers record how the student complied with the preliminary admission requirements for higher education (CREHE Preliminary). Compared to the ERR the information goes back a longer period in time as data collection started in the early 1980s.

Quite new are the Education Number Registers (ENR) for secondary education, including higher secondary vocational education and adult education programmes. The ENR focus on all grades a student passes through, and it is easy to find the corresponding SCED or ISCED codes. The ENR for secondary general and lower secondary vocational education started in 2002/'03; for higher secondary vocational and secondary adult education in 2004/'05. In the beginning these registers did not yet manage a hundred percent coverage of the student population.

One shortcoming of the registers on secondary education is that they are restricted to publicly financed educational institutes. The information for students in private schools[8] is incomplete. These registers also have no information on people who studied outside the country.

3.2.3. Higher or tertiary education

The Central Register for Enrolment in Higher Education (CREHE) registers information on a yearly basis about students at the university level (from 1983) and in higher vocational education (from 1986[9]). This includes data on certificates. Because of its long registration period CREHE can now present data on an entire generation aged 18-40 who were enrolled in university or in other programmes of higher education as long as they were not in a foreign country or at privately financed institutions[10]. Here too SCED and ISCED codes can be easily derived from the CREHE education codes.

3.2.4. Other administrative registers

Beside the education registers mentioned above there are some administrative registers which can be useful when supplementary data on education levels are needed.

The CWI register was originally thought to be a promising source for information on education levels, owing to the large scale of its target population. CWI is the Centre for Work and Income, the public employment service. When new clients are registered, data on their education level are recorded, and changes are updated as long as they stay clients. Between 1990 and 2007 information has been collected on more than 5 million people.
Unfortunately the educational attainment registration in the CWI register is of insufficient quality (Bakker et al., 2008). What is registered by CWI seems to have more to do with competencies for the labour market than with schooling. There are indications for that from micro-linkage with other sources, such as the LFS. To gain insight into the differences there is research going on. Nevertheless, when estimating the distribution of education levels the administrative variable educational attainment in the CWI register can be useful as an auxiliary variable. In spite of the differences with the desired statistical concept of educational attainment, there is undeniably a high correlation between both. See also section 5 on weighting matters.

[8] A rough LFS-based estimate indicates that some 5 percent of the student population of 2005 got secondary or higher education (of more than 12 months) that was not publicly financed. For the near future the intention is to add data on private secondary education to the ENR.
[9] In the initial two years there was some underreporting.
[10] See footnote 8.

Another inferior point of the CWI register is that the information is less detailed, only five different education levels are distinguished.

Another administrative register is the Student Finance Register (SFR) with information on study grants from the government. In the Netherlands most students in higher education, and also students over eighteen in upper-level secondary vocational education, receive such a grant during a few years. It could be a rich source, with numerous students registered from 1995, but unfortunately there are no data on certificates and the information on study stages suffers lack of detail. So it is of limited use for deriving educational attainment.
Then, there is a financial register from 2001 onwards, which registers students in secondary education for whom school fees have to be paid by law. Originally, the target group was students over 16, but from 2006 it is restricted to student population over 18. The problem with this Register of School Fees (RSF) is that we do not know to what type of secondary education the school fees refer, nor is there any information available about certificates. Therefore the RSF cannot to be used for the purpose of deriving education levels.

There is no register with information on the highest attained education level for older people. To get some more insight in their education levels it has often been suggested to make use of the last traditional population census of 1971. That is to say, as far as these persons were at the end of their education career in or before 1971. Unfortunately, linking data of the 1971 Census to present-day files is difficult due to a lack of appropriate personal identifiers in those days. So, we have to forget the idea of using the 1971 Census.

4. Micro-integration

4.1. From education sources to Education Archive

After micro-linkage with the Population Register, all students of the sources mentioned in the preceding section are supplied with a unique identifier Record Identification Number for the person (RIN-person). This means that the information of these sources can be brought together and sorted on personal level in a huge database, the so-called Education Archive. This archive shows the education careers of the total population as extracted from the sources. It is updated every year with new information on education programmes. In other words, the Education Archive 2007 stores all available information on education careers up to 2007 in cumulative form, resulting in almost 65 million records. The information in the Education Archive can be conflicting at times, when data on the same subject come from different sources. Figure 2 gives an example of a fragment of the Education Archive 2007.

Figure 2. Education Archive 2007, fragment (fictitious example).

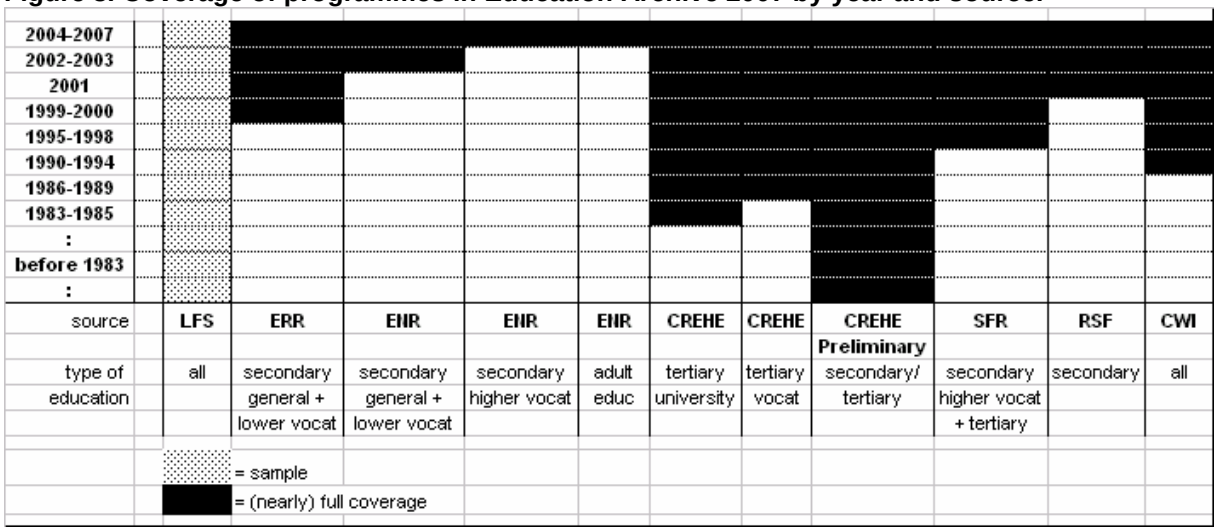
Nr	RIN-person	Sex	Birthdate	Source	Type of Education	Start date	End date	Certificate	SCED-code	sample weight	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
1502377	R006274060	M	19850202	ERR'01	Lower Secondary Vocational	----------	20010625	yes	012849	1,000	.....
1502378	R006274060	M	19850202	RSF'01/'02	----------	20010801	20020731	-	------	1,000	.....
1502379	R006274060	M	19850202	RSF'02/'03	----------	20020801	20030731	-	------	1,000	.....
1502380	R006274060	M	19850202	CREHE'07	Higher Vocational bachelor	20050901	20060831	no	001117	1,000	.....
1502381	R006274060	M	19850202	CREHE'07	Higher Vocational bachelor	20060901	20070831	no	001117	1,000	.....
1502382	R006274060	M	19850202	SFR'05	Higher Vocational Education	20050901	20051231	-	------	1,000	.....
1502383	R006274060	M	19850202	SFR'06	Higher Vocational Education	20060101	20061231	-	------	1,000	.....
1502384	R006274060	M	19850202	SFR'07	Higher Vocational Education	20070101	20071231	-	------	1,000	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
3573951	R046879435	M	19590523	CWI'94	Higher Secondary Education	----------	----------	-	------	1,000	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
9929931	R165529489	F	19811215	LFS'96	Primary Education	----------	19930630	yes	800001	97,168	.....
9929932	R165529489	F	19811215	LFS'96	Secondary general	19930701	----------	no	002538	97,168	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

The information from the sources is kept as authentic as possible, as ought to be the case in an archive. So, micro integration at this stage of the process is restricted to correcting obvious errors and a bit of harmonization. An example of the latter is the standardization of internal education codes of the different sources to 6-digit SCED codes. At the end of the process the integrated archive gets its place in the SSD core.
For RIN-person R006274060 in figure 2 the only register information on lower secondary vocational education is from ERR (exam results). After passing his exam at 16, he must have attended secondary education for two years, as the RSF shows. This was probably higher secondary vocational education, but it cannot be confirmed as there were no registers on that sort of education until 2004.

There are no observations for the period between September 2003 and September 2005. He may have worked or gone abroad to go to college for two years. There is no register information on studies outside the country. It is clear from this example that it cannot be guaranteed that the information on someone's education career in the Education Archive is always complete.
RIN-person R046879435 looked for a job in 1994, according to CWI'94. He was registered as someone with higher secondary education. There are no further details on his education.
RIN-person R165529489, who was interviewed for the LFS'96, was evidently attending secondary general education from 1993 to 1996, but one does not know what happened to her afterwards, because there were no ENR registers by that time. Did she pass her exam in '96/'97 or '97/'98? That will remain the question, as the ERR did not exist yet.

Figure 3 outlines the coverage in the Education Archive 2007 by year and source. Only LFS and CREHE Preliminary cover the period before 1983. CREHE Preliminary covers it only for those who attended higher education in or after 1983.

Figure 3. Coverage of programmes in Education Archive 2007 by year and source.



4.2. From Education Archive to Educational Attainment File

The ultimate goal is to build an Educational Attainment File containing the highest attained education level of an individual with and without certificate. Education levels in the Educational Attainment File are presented as 2-digit SCED-level codes (Schaart et al., 2008), see figure 4. For Census purposes ISCED-level codes can be added.

Figure 4. SCED2006 and corresponding ISCED1997- level codes.

SCED2006 level code	Description SCED2006 level code	ISCED1997 level code
10/20	Not more than primary education	0 / 1
30	Secondary education, 1st stage	2
31	Secondary education, 1st stage, lower level	2
32	Secondary education, 1st stage, intermediate level	2
33	Secondary education, 1st stage, higher level	2 / 3
41	Secondary education, 2nd stage, lower level	3
42	Secondary education, 2nd stage, intermediate level	3
43	Secondary education, 2nd stage, higher level	3 / 4 / 5
51	Higher education, 1st stage, lower level	5
52	Higher education, 1st stage, intermediate level	5
53	Higher education, 1st stage, higher level	5
60/70	Higher education, 2nd and 3rd stage	5 / 6

Description ISCED1997 level code
0. Pre-Primary education
1. Primary education
2. Lower secondary education
3. Upper secondary education
4. Post secondary non-tertiary education
5. First stage of tertiary education
6. Second stage of tertiary education

Source: Schaart et al. (2008).

To derive someone's educational attainment, data have to be selected from the Education Archive. Now data in the Education Archive are stored practically in their original state, depending on the source they come from. Therefore differences in definition, level of detail, quality, measurement and

coverage are inevitable. In order to get an unambiguous presentation of educational attainment, micro-integration has to be performed in successive stages.

4.2.1. Selection from the Education Archive

The first step to get education levels at the reference date is to select all records in the Education Archive representing education careers until that date. Before this can be done, it may in some cases be necessary to derive or impute missing start and end dates of an education programme. This is the micro-integration step of imputation or derivation for missing values.

Second, only those data in the Education Archive are selected that belong to the target population of the Education Attainment File. That is, the population in the Population Register (PR) at the reference date. The Education Archive also has information on foreign students, but they are not part of the target population. In this second stage of micro-integration the differences in definitions between source and target population are adjusted.
In addition, records are added from the PR for boys and girls under 12 at the reference date, since there is no source on Primary Education. They are classified with education level 'not more than primary education' (SCED2006-level 10/20).
Apart from that, the 12 to 14 year olds from the PR who are not represented in the Education Archive are included. The idea behind this is that although they are missing in the Education Archive, attendance is compulsory for these ages. Their absence in the Education Archive probably has to do with an omission in the administrative registers. For this particular population the highest education level with certificate will be 'not more than primary education' (SCED2006-level 10/20). The highest education level without certificate will be determined as 'secondary education first stage' (SCED2006-level 30).

Third, not all source information in the Education Archive is useful for deriving educational attainment. RSF records have no information on education levels and will not be selected. SFR records on the other hand will only be selected if the SFR education codes have sufficient detail, which is not always the case. The information on education levels in CWI records also suffers lack of detail, and so these data will not be used to determine the highest education level. However, the CWI data are going to perform an important role in the weighting strategy, as will be set out in section 5. Micro-integration here is a matter of quality assessment of the sources, and a decision is made which sources to use.

4.2.2. Determination of education levels at the reference date

In order to be able to derive someone's educational attainment level, one has to compare education levels of the different programmes in the individual's education career. Deriving comparable levels is in fact a micro-integration activity of harmonization. It is possible to find a corresponding SCED level code for each 6-digit SCED code. SCED levels, however, cannot be transformed into ISCED levels, as there is no 1-on-1 relationship between SCED2006 and ISCED1997 levels (see figure 4). So, ISCED level codes should also be determined with 6-digit SCED-codes as the starting point.

Determining educational attainment would now simply seem to be a matter of locating the highest attained education level (with and without certificate) of all programmes. However, it is not that simple. As we have seen in subsection 4.1 the coverage of education careers in the Education Archive is to some extent fragmented. Looking back to figure 2 we cannot be certain about the educational attainment of RIN-person R006274060 in 2005, because of an information gap for school-years 2003/'04 and 2004/'05. And what about the educational attainment of RIN-person R165529489 in 1997 of 1998? Did it change after she was interviewed for the LFS in 1996? The information may be out-of-date. It is clear that before education levels can be compared at reference date their validity has to be assessed for that date. In terms of micro-integration we can interpret the problem as one of measurement errors that must be corrected for. This will be tackled in the next subsection.

4.2.3. Assessing the validity of education levels at the reference date

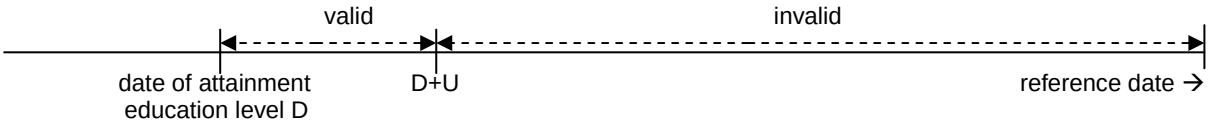
A decision strategy has been developed to assess if the information on someone's education level is still valid at reference date. Various deterministic and stochastic decision rules are distinguished.
An example of a deterministic decision rule applied is the following. If a record in the Education Archive originates from the ENR register, it will be assessed valid at reference date no matter from which year the source information dates. The argumentation behind this is that if someone went on to

a higher education level it would be in the Education Archive, because nearly all possible continuation programmes can be found in the archive. One exception is that the person may have continued in private education, which is not covered yet by registers in the Education Archive. In that case the validity may be assessed wrongly.
A trivial deterministic decision rule is the one in which all records with SCED-level code 60/70 are assessed valid at reference date. The simple reason is that there is no higher level attainable.
Another trivial deterministic decision rule is that education levels in the Labour Force Survey (LFS) should automatically be considered valid if the LFS-interview date lies beyond the reference date for which educational attainment has to be determined. This is because the complete education career until interview date is registered in the LFS.

The decision rules can validate different education levels simultaneously! Think of the examples mentioned above: an ENR record with say SCED-level code 43 and a record of the same person with SCED-level code 60/70. Of course, it is impossible for both to be valid at the reference date. However, in the end educational attainment will be determined as the highest level declared valid. In other words, the lower levels will be overruled.

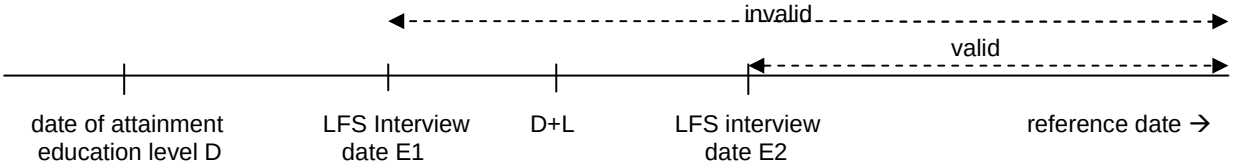
Stochastic decision rules are based on the probability that an education level is still valid at reference date. This is done with the help of non-parametric survival analysis models (Life Tables method)[11].
The first stochastic decision rule judges until when an education level is still valid after date of attainment D. With the Life Tables method an upper bound U is determined for the number of years after attainment for which the probability is more than 95 percent that this education level has remained unchanged. The bounds are dependent on age, background (Dutch, other Western or non-Western origin), and education level. See figure 5.

Figure 5. Stochastic decision rule 1



The second stochastic decision rule only applies to LFS records. It is based on the idea that if someone attained a certain education level a long time ago the probability is very high that this level remains unchanged. With the Life Tables method a lower bound L is determined for the number of years between date of attainment D of the education level and the LFS interview date E. See figure 6. If E-D ≥ L the probability is more than 95 percent that educational attainment will not change anymore. As with the first stochastic decision rule the bounds are dependent on age, background and level of education.

Figure 6. Stochastic decision rule 2



With an interview date following the attainment date as soon as E1 there is not enough certainty that the highest education level will not rise after E1. If the interview date is much later, say E2, then the probability is 95 percent or more that the highest education level remains the same after E2.

There is also a third stochastic decision rule, again only for LFS records, which is some sort of a generalization of the first two rules.

[11] The survival function S(t) = P[T ≥ t] gives the probability that the educational attainment has not changed within t years. The distribution was determined empirically on the basis of the LFS for a number of years.

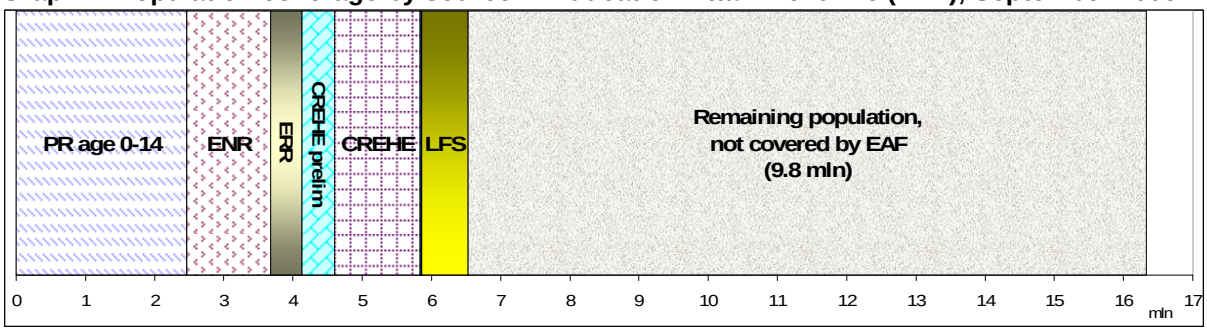
4.2.4. Determining educational attainment at the reference date

After the validity of all education levels has been assessed it is now possible to select the highest of all levels declared valid at reference date and to fill the Educational Attainment File. In principle it does not matter which source supplied the highest level. There is one exception on this. The rule is that if both LFS and a register contribute information on the highest level, it will be taken from the register. This distinction is relevant because LFS records get a sample weight, while register records do not. The reason for preferring register information is because registers are thought to be more accurate.

The Educational Attainment File is provided with two types of educational attainment. For the first type it is essential that someone is qualified and certified at the highest level. For the second type it is sufficient to be educated at the highest level without necessarily being certified. For students who have not passed an exam yet the two educational attainment types will generally differ.

Special attention should be paid to downgrading. Suppose there are two records in the Education Archive representing a part of someone‘s education career. The first record of this person refers to secondary education SCED-level code 33, with certificate. The second describes this person attending university. It would be wrong to qualify this person during his university period with educational attainment SCED-level code 33 (with certificate). After all, a higher preliminary education level is required for university admission. For some reason the Education Archive lacks the information on an intermediate education programme. In such cases the method of downgrading is applied, which implies that the highest level with certificate is determined as the minimum (downgrade) level that is required for attending the education programme concerned. In the example of the university student this would be SCED-level code 43.

Graph 1. Population coverage by source in Education Attainment File (EAF), September 2005



In the end all valid educational attainment records are stored in an Educational Attainment File (EAF). For reference date September 2005 this file contains in total 6.5 million records, of which 5.8 million records from registers and 675 thousand records from LFS. See graph 1. The contribution of SFR is so small that it is imperceptible in the graph. In terms of integration, population differences remain between observed and total population (16.3 million individuals). This means that a large gap of 9.8 million, or 60 percent, still has to be bridged. How this happens is discussed in the next section.

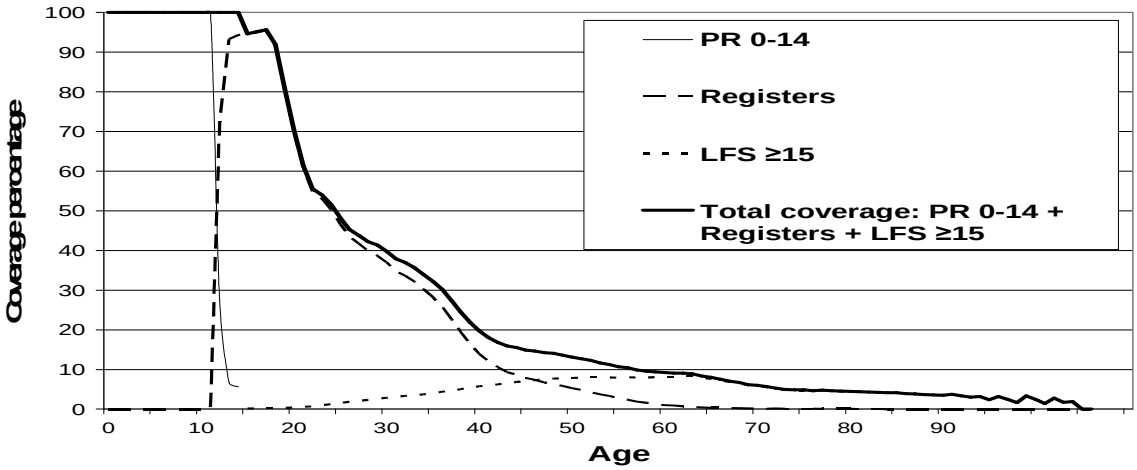
5. Weighting Strategy

The Education Attainment File (EAF) consists of a mixture of register and sample records. For 2005 the EAF covers about 40 percent of total population. However, the coverage is not uniformly spread over population, as graph 2 shows. It is not surprising that younger people are better represented by these education registers since the registers have not existed very long.
The EAF has full coverage of children aged 0-14 at the reference date, as we have seen in section 4.2.1[12]. From the age of 15 we start with a coverage of 95 percent, predominantly by registers (see graph 2).
As age rises the register coverage curve shows a sharp decline in the first section, from 81% coverage for age 19 to 61% for 21 year olds. From then on the decline is less steep. The coverage of

[12] In age 0-11 the Education Attainment File records only come from the Population Register (PR). From ages 12 to 14 records come either from registers or the PR. There is not much response from this group in the LFS, so we decided not to use this source at all for these ages.

the register part drops to a little more than 10 percent for the ages of forty. The curve for registers shows a gradual decrease to less than 1 percent when people are over sixty.
The LFS coverage curve starts very modestly with a contribution of no more than 2 percent up to age 30. This is, of course, because this area is strongly represented by the registers[13]. The curve shows a gradual incline up to 8 percent for the age of fifty. For people over sixty the LFS curve almost coincides with the curve which represents total coverage, in other words LFS is almost entirely responsible for coverage in that age class. At the same time the LFS curve shows a decline for these ages. This is mainly because there is less sampling in the LFS amongst people over 65. The undersampling is compensated by giving them a higher final LFS weight.

Graph 2. Population coverage by source and age in Education Attainment File, September 2005



Since older ages are underrepresented in the EAF, this file is not representative for total population. The EAF will also not always be representative for other population characteristics, because of the relatively great share of register information. These registers were set up for a specific target population, after all. Another reason why they are not representative is the selective loss which arises, when records are declared invalid by the different decision rules mentioned in section 4.2.3.

A weighting strategy is applied to make the EAF representative. This is done as follows.
The register records (5.8 million in 2005) are an integral representation of the register‘s target population. For example, the CREHE describes the population in higher education at publicly financed institutes. Therefore a final weight equal to 1 is attached to the register records, which in fact is the same as counting them. So in 2005, a total of 5.8 million register records got a weight equal to 1.
What remains is the LFS (675 thousand records in 2005). This has to be reweighted to bridge the gap (9.8 + 0.675 million in 2005) between register population and total population. For 2005 this implies that an average calibrated LFS weight of 15.5 (9.8+0.675 divided by 0.675) is needed to cover the remaining population (LFS inclusive). As the LFS has its own specific sampling design it would be wise to use the LFS weights as initial weights. These, for their part, have to be reweighted[14] in order to get a representative outcome.

A complicating factor is that there are several LFS involved from different years. For EAF 2005, for example, eleven LFS (1996,..,2006) have been used. Respondents generally only participate once[15] in the LFS, so when the aim is to collect as much data as possible, it would be wise to make use of all the available LFS. Of course, not all LFS observations will be accepted for the EAF. For example, when LFS of earlier years are used and the EAF reference date lies far ahead, fewer observations are declared valid if stochastic decision rule 1 is applied.
Each LFS has its own year-specific sampling design and sample size. One would like to prevent that LFS has a more dominant contribution for some years than for other years. To realize this a **scaling** procedure is applied to the LFS-weights so that: 1) the total sum of all the scaled weights is equal to

[13] As mentioned in section 4.2.4 registers get priority if there are observations in both registers and LFS.
[14] Register records are not affected by reweighting. Register weights remain equal to 1.
[15] In fact a LFS-respondent is approached five times during the year in which he participates.

the remaining population including LFS (9.8+0.675 million in 2005); 2) the average scaled weight per year is the same for each year[16].

Together with the register records the scaled LFS weights add up to the total population. So now we have complete coverage of total population, however, the weighted sums for subgroups will still differ from the population margins. Therefore, apart from scaling, we have to take one more step, which is **calibration** (reweighting) of the scaled weights. Besides consistency with population margins, the function of calibration is also to lower variances for the estimates of the target variable, and to reduce selective non-response bias.
For the EAF of 2005 several auxiliary variables were used in the weighting model[17]:

1) demographic variables such as gender x age (mainly 5-year classes), marital status, countries of origin (7 categories plus distinction first/second generation) and region;
2) socio-economic variables such as employee (yes/no), self-employed (yes/no), other labour income (yes/no), unemployment benefits (yes/no), disablement benefits (yes/no), social assistance benefits (yes/no), pension/life insurance benefits (yes/no), other benefits, socio-economic category, and income (20%-groups);
3) educational attainment in the CWI register is added as an auxiliary variable. As mentioned in section 3.2.4 the CWI version of educational attainment was rejected for usage as target variable, but it was considered suitable as an auxiliary variable in a weighting model because of the high correlation with the desired concept of educational attainment.

Once scaling and calibration are carried out, one can get representative estimates of the distribution of the educational attainment level for (sub)populations. Table 1 shows educational attainment for people in the Netherlands aged over 15 in 2001 and 2005. Three basic categories are distinguished, lower education (ISCED 0/1/2), intermediate education (ISCED 3/4) and higher or tertiary education (ISCED 5/6). There is not much difference between the Census and LFS percentages for 2001, nor between the EAF and LFS percentages of 2005. The table suggests that the level of education in the Netherlands has on average gone up between 2001 and 2005.

Table 1. Education levels of population aged 15+ in the Netherlands (in %), 2001 and 2005.

Education level	Census 2001[1)]	LFS 2001	EAF 2005 [2)]	LFS 2005
ISCED 0/1/2 Lower Education	42.9%	42.5%	38.7%	38.0%
ISCED 3/4 Intermediate Education	39.0%	38.4%	38.5%	38.9%
ISCED 5/6 Tertiary Education	18.2%	18.7%	22.7%	22.5%
Unknown		0.4%		0.6%
Total population age 15+	100.0%	100.0%	100.0%	100.0%

[1)] based on LFS 2000 and 2001. Reference date 1 January 2001.
[2)] based on administrative education registers and LFS. Reference date September 2005.

The weighting model for EAF 2005 may have been too generously specified with auxiliary variables. This idea came when we found some unexpected and implausible outcomes for some small

[16] Scaling procedure in formulae, worked out for LFS(1996,..,2006):
$w_{t,i}$ = LFS-weight of observation i=1,..,n(t) in LFS of year t; N_{remain} is remaining population (LFS inclusive) to be covered by LFS; n_t is sample size of LFS in year t; t=year 1996,..,2006).

With scaling factor $\lambda_t = (n_t / \Sigma_{t=1996,..,2006}\, n_t)\ .\ (N_{remain} / \Sigma_{i=1,..,n(t)}\, w_{t,i})$ we get the combined result that:
1) the scaled weights $\lambda_t.w_{t,i}$ altogether add up to N_{remain} : $\Sigma_t\Sigma_i\ \lambda_t.w_{t,i} = N_{remain}$;
2) the average scaled weights $\Sigma_{i=1,..,n(t)}\ \lambda_t.w_{t,i} / n_t$ for each year t are identical: $N_{remain} / \Sigma_{t=1996,..,2006}\ n_t$.

[17] For a solution of the weighting model a method of linear weighting is applied.

$$\min_{w} \sum_{i=1}^{n} \left(\frac{w_i}{d_i} - 1 \right)^2 \text{ subject to } X_s' w = x_{pop} \text{ (calibration equations)}$$

w_i = final calibration weight; d_i = scaled LFS-weight (initial weight for calibration procedure); w is vector of w_i; X_s is matrix representing subpopulations related to the auxiliary weighting variables and x_{pop} the vector of corresponding population margins.
The solution is found with BASCULA, a tool for weighting sample survey data (reference Bascula 4).

subpopulations. One of the objectives of the weighting model was to achieve consistency with as many population margins as possible. However, the danger of an abundant model is that the final weights may be troubled too much by fluctuation. This can have a disrupting effect on the accuracy of the target variable (educational attainment in our case) in cells with few observations. Nascimento Silva & Skinner (1997) point out in a simulation study that adding auxiliary variables in a regression model causes the variance of the regression estimator to drop initially, but by adding variables the variance will tend to increase from a certain point on.
For this reason we now consider revising the weighting strategy. Although still tentatively, the outcome may be that the auxiliary information will have to be restricted to variables that show strong coherence with educational attainment. If so, a search strategy must be applied to find an optimal weighting model with only relevant auxiliary variables.

If in addition to such a parsimonious model, consistency of other than the weighting variables is still considered important, one could consider applying the method of **repeated weighting**. This method was also used for the Virtual Census 2001, see Linder (2004). Repeated weighting is in fact based on the repeated application of the regression method, and it is not the same as calibration (reweighting). Calibration results in a fixed set of survey weights for the sample concerned, whereas with repeated weighting a new set of weights (based on the survey weights) is derived per table in order to get consistency with population margins or other tables estimated before.

6. Accuracy

Statisticians generally want to know if the results of their estimates are reliable. When samples are used instead of registers one is faced with sampling errors. The standard literature on sampling theory mostly refers to estimation on sample data.
There is less literature in the case of combined register and sample data. It is intuitively clear that a higher share of register data reduces the effect of sampling errors. But even then, these errors are still there, so we need statistical measures to determine whether (sub)population estimates are accurate enough.

In the Virtual Census of 2001 a rule of thumb was to allow publication of table cells if they were based on at least 25 sample observations. The threshold of 25 observations corresponded to an estimated coefficient of variation (CV) of at most 20 percent, which was considered an acceptable tolerance. The threshold formula is described in Knottnerus (2009)[18]. If a table cell has sample data as well as register data, the threshold will also depend on the number of register observations in the cell. The more register observations, the less sample observations are needed to keep the accuracy level. Knottnerus generalized the threshold formula for the case of combined register and sample data. The formula only holds when the number of sample observations is at least 25[19]. This is unfortunate because we are particularly interested in the accuracy for smaller subpopulations, also if sample size

[18] Suppose N is population size; n is sample size; N(g) is total population in cell g; n(g) is number of sample observations in cell g; p=N(g)/N; q=1-p and f=n/N.

With n large enough, the variance of $\check{N}(g)$ can be approximated as: $var(\check{N}(g)) = N^2pq(1-f)/n$

With the assumption that the average sample fraction f is very small (e.g. LFS sample fraction is about 1 percent), and p is very small (i.e. relative small subpopulation) the coefficient of variation (CV) can be approximated as $[q(1-f)/np]^{½} \approx [1/n(g)]^{½}$. So, a CV ≤ 20% implies n(g) ≥ 25.

[19] Suppose we have sample observations as well as register observations in cell g; $N_1(g)$ is the number of register observations in cell g; $N_2(g)$ is the weighted number of sample observations in cell g; $N(g)= N_1(g)+ N_2(g)$; $n_2(g)$ is the (unweighted) number of sample observations in cell g.

For a coefficient of variation of not more than A it is required that $n_2(g) \geq (1/A)^2 . [N_2(g) / (N(g)]^2$.

With a higher $N_1(g)$ the threshold decreases. The problem with the formula is that, because approximations are applied, it is only valid with sufficient $n_2(g)$. Knottnerus suggests that $n_2(g)$ should be at least 25 for the formula to be valid. See Knottnerus (2009).

drops below 25. The problem could be tackled by deriving exact formulae for variance estimators, but that is a very difficult project in the case of combined register and sample survey data.
An alternative approach of the problem was applied by three methodologists of Statistics Netherlands. Kuijvenhoven, Scholtus and Schouten (2009) used a method of bootstrap resampling to make estimates for accuracy measures. Their ideas were based on Canty and Davison (1999). The bootstrap method as applied by Kuijvenhoven, Scholtus and Schouten can be summarized as follows. The three methodologists inflated the sample part of the Education Attainment File (EAF) up to the entire population, by adding copies of each sample record. Then they took a number of so-called bootstrap samples without replacement from the population file so constructed, with each sample as big as the original sample part in the EAF. Kuijvenhoven, Scholtus and Schouten found that 500 bootstrap replications were sufficient for convergence of the bootstrap variance[20] [21].

Table 2 compares the accuracy of the educational attainment estimates from the LFS and from the EAF for an example of small subpopulations. In this case the example is young people of Turkish origin with an education level equivalent to a master's degree or higher. The LFS 2005 does not have enough observations (less than 25) to give accurate information (coefficient of variation (CV) over 20 percent). With the EAF 2005 the number of sample observations is also under 25 in all cells, but the results are sufficiently accurate for the age group 25-30 (CV of at most 20 percent).

Table 2. Coefficient of variation (CV) for some highly educated (master's degree level or higher) Turkish males and females, aged 18-30, 2005.

	LFS 2005			EAF 2005				
Subpopulation	n(g)	$\check{N}(g)$	CV [1)]	$N_1(g)$	$n_2(g)$	$\check{N}_2(g)$	$\check{N}(g)$	CV [2)]
Turkish males; age 18-24	0	0	n.a.	39	1	129	168	32%
Turkish females; age 18-24	2	382	>20%	58	3	157	215	42%
Turkish males; age 25-30	7	1.175	>20%	286	9	415	701	20%
Turkish females; age 25-30	6	1.071	>20%	321	9	367	688	19%

n(g) is number of observations in cell g in LFS 2005; $\check{N}(g)$ is estimate of total population in cell g; $N_1(g)$ is number of register observations in cell g in EAF 2005; $n_2(g)$ is number of sample observations in cell g in EAF 2005; $\check{N}_2(g)$ is weighted number of sample observations in cell g in EAF 2005; n.a. is not applicable
[1)] approximation (see footnote 18); [2)] bootstrap estimation.

7. Concluding remarks

The introduction of administrative sources on education has enabled the creation of an Education Attainment File for the population of the Netherlands. The results so far are promising, and offer a serious opportunity for innovation in the Census of 2011 of its method to produce educational attainment figures. With the new approach of combining register and sample data, we expect to be able to publish more detailed table cells. Bootstrap variance estimation makes it possible to determine whether the estimate of education levels is accurate enough for dissemination. If consistency requirements within the census table programme are to be met, the bootstrap method described in this paper will have to be worked out for that purpose.
An elaborate Dutch-language survey of the methodology of constructing educational attainment from administrative sources and sample surveys can be found in Linder & Van Roon (2008).

[20] The following formula was used as variance estimator for cell g with bootstrap samples b=1,..,B:

$Var_B(g) = (1/(B-1)).\sum_b [\ \check{N}_b(g) - 1/B.\sum_b \check{N}_b(g)\ ]^2$.

Only for sample sizes less than 5, a more sophisticated version of this estimator was applied, as solution for the possibility that nobody from cell g was found in one or more bootstrap samples.

[21] For estimation of confidence intervals it is suggested by literature on bootstrap methods to increase the number of bootstrap samples to at least thousand.

References

1. Bakker, Bart F.M., Frank Linder & Dominique van Roon, *Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys,* paper presented at IAOS Conference on Reshaping Official Statistics, Shanghai, 2008.
2. Bascula 4, *A tool for weighting sample survey data and variance estimation*, http://www.cbs.nl/en-gb/menu/informatie/onderzoekers/blaise-software/blaise-voor-windows/productinformatie/bascula 4, Statistics Netherlands.
3. Becker, G.S., *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* Third edition, University of Chicago Press, Chicago, 1993.
4. Canty, A.J. and Davison, A.C., *Resampling-based variance estimation for labour force surveys',* The Statistician volume 48, pp. 379-391, 1999.
5. CEC (Commission of the European Communities), *Communication from the Commission. „Education & Training 2010" The succes of the Lisbon Strategy hinges on urgent reforms. Draft joint interim report of the detailed work programme on the follow-up of the objectives of education and training systems in Europe,* 2003.
6. CEC (Commission of the European Communities), *Commission regulation implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns*, 2009 (draft).
7. Daas, P., S. Ossen and J. Arends-Tóth, *Framework of Quality Assurance for Administrative Data Sources*, paper presented at the International Statistical Institute Conference 2009 in Durban, South-Africa, 2009.
8. Knottnerus, P., *Remarks/discussion with respect to LFS-margins for small subpopulations and with respect to LFS weighting models,* internal note, Statistics Netherlands, 2009.
9. Kuijvenhoven, L., S. Scholtus and B. Schouten, *Construction of accuracy measures for the Education Attainment Files* (in Dutch), Statistics Netherlands, 2009.
10. Laan, P. Van der, *Integrating Administrative Registers and Household Surveys,* In Netherlands Official Statistics, volume 15 (Summer 2000): Special Issue, pp.7-15, 2000.
11. Linder, Frank, *The Dutch Virtual Census 2001: A new approach by combining Administrative Registers and Household Sample Surveys,* In Austrian Journal of Statistics, volume 33, pp. 69-88, 2004.
12. Linder, Frank and Dominique van Roon, *Methodology construction educational attainment files* (in Dutch), Statistics Netherlands, 2008.
13. Mincer, J., *Schooling, experience, and earnings,* Columbia University Press, New York, 1974.
14. Nascimento Silva, P.L.D. and C.J. Skinner, *Variable selection for regression estimation in finite populations*, Survey Methodology, volume 23 no.1, pp. 23-32, 1997.
15. Queen's speech, in Dutch, http://www.regering.nl/Het_kabinet/Troonrede_2007, 2007.
16. Queen's speech, in Dutch, http://www.regering.nl/Het_kabinet/Troonrede_2008, 2008.
17. Schulte Nordholt, E. & F. Linder, *Record matching for census purposes in the Netherlands,* In: Statistical Journal of the IAOS, volume 24, pp. 163-171, 2007.
18. Schaart, Roel, Sue Westerman and Mies Bernelot Moens, *The Dutch standard classification of education SOI 2006,* Statistics Netherlands, spring 2008.
19. UN, *Principles and Recommendations for Population and Housing Censuses, Revision 2,* 2008.
20. UNECE, *Conference of European Statisticians Recommendations for the 2010 Censuses of Populations and Housing, prepared in coöperation with the Statistical Office of the European Communities (EUROSTAT),* United Nations, 2006.
21. UNECE, Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics, 2007.