IFAU
Institute for Evaluation of Labour
Market and Education Policy

On the use of register data in educational science research

Erik Mellander

WORKING PAPER 2016:22

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

On the use of register data in educational science research[a]

by

Erik Mellander[b]

November 27, 2016

Abstract

Register data are described, in general terms and in specific terms, focusing on informational content from an educational science perspective. Arguments are provided for why educational scientists can benefit from register data. It is concluded that register data contain lots of information relevant for educational science. Furthermore, two specific features of register data are considered: their *panel data nature*, implying that register data analyses under certain conditions can account for aspects on which the registers are not informative, and that they contain *intergenerational links*, facilitating the separation of genetic and environmental influences on learning. It is observed that while register data do not contain direct links between students and teachers this shortcoming can be overcome by merging register data with survey data on these links. Being population data, register data enable analyses which are not feasible to conduct by means of survey data. An illustration is provided of how quantitative and qualitative researchers can benefit from combining register-based statistical analyses with in-depth case studies. The use of register data in evaluations of causal effects of educational interventions is also described. To facilitate the exploitation of the aforementioned advantages, a discussion of how to get access to register data is included.

Keywords: Register data, Nordic, panel data, intergenerational links, ethical review, combining quantitative and qualitative methods, causal effect evaluations

[a] I'm grateful to Lena Tibell for encouraging me to teach on this topic; my lecture notes were the main input to the paper. Helpful comments from Mary James, Caroline Hall and Sara Martinson on a previous version of the paper are gratefully acknowledged.
[b] Institute for Evaluation of Labour Market and Education Policy (IFAU). Email: erik.mellander@ifau.uu.se.

Table of contents

1 Introduction

Register data is fundamentally a Nordic phenomenon. Similar data, providing information about a country's entire population in a large number of different respects, are not available elsewhere.[1] With respect to research, register data provide the Nordic countries with a formidable advantage compared to the rest of the world. However, while register data for a long time have been extensively used in medical science and epidemiology, as well as in many social sciences, e.g., sociology, economics and political science, their potential has, as of yet, been little exploited in educational science. Partly this may be due to comparatively less emphasis on quantitative skills in the PhD program in educational science; knowledge of quantitative methods does facilitate the use of register data. Presumably, there is also still a lack of information about what empirical educational scientists stand to benefit from using register data; cf. Vetenskapsrådet (2015).

A discussion of the research training offered to graduate students in educational science is beyond the scope of this article. Its more modest aim is to address the second explanation: to reduce the informational problem by discussing, from an educational science perspective, five advantages of register data.

The first advantage is that register data are very rich and detailed, both across individuals and organizations, and over time. In addition, two very useful properties of register data are that they constitute *panel data* and contain *intergenerational links*. The panel data structure means that there are repeated observations on individuals and organizations. The presence of intergenerational links implies, i.a., that a student's parents and siblings can be identified. See further Section 4.

The second advantage is that register data enable powerful combinations of quantitative and qualitative research methods. This is illustrated in Section 6

The third advantage has to do with the fact that the analysis of some research problems requires data that are either very expensive or impossible to collect by means of surveys but which are readily available in register data. An example is provided in Section 7.

[1] In some countries outside the Nordic region there are register data containing information about specific fields like, e.g., health and education, and on specific issues. One example is England where there are data for all students attending state schools on results in national achievement tests, cf. Crawford et al. (2010). However, unlike in the Nordic countries, this register cannot be linked to other registers. For instance, it cannot be linked to information about the students' parents to enable analyses of possible influences of family background on test results.

The fourth advantage is that by combining register data with survey data it is possible to construct data sets with properties which cannot be achieved by the compilation of either survey data set or register data sets alone. This advantage is illustrated in Section 8.

The fifth advantage is that register data under certain conditions enable large-scale evaluations of causal effects of educational policy interventions; cf. Section 9.

Preceding detailed discussions of the advantages of register data just mentioned, the next two sections consider, respectively, typical features of register data and different ways to make use of register data, again from an educational science perspective.

2 Typical features of register data

The most salient feature of register data is that they are *population data*, defined and measured at *very low levels of aggregation*. Specifically, the data concern individuals, residences, workplaces (like, e.g., schools) and small geographical units. For example, consider persons that are residing in a Nordic country and are enrolled in upper secondary school: there are register data about all of these individuals.

That register data are population data is, in itself, a huge advantage, because it means that register data can be matched with an arbitrary other data set without any observations in that other data set being lost. This is quite different from, say, when two survey data sets are being matched; in the latter case, the resulting matched data set will only contain (complete) information for the intersection of the two original data sets.

Another common property of most types of register data is that, typically, they have not been compiled to support research but for administrative purposes, like tax collection, society planning (of education, for instance), censuses, and health documentation, etc. In consequence, register data are seldom ideally structured from a researcher’s point of view. However, in most cases this is not a big problem; today there is plenty of good and inexpensive software that efficiently can handle many kinds of data sets.

There are exceptions to the rule that register data sets have been compiled for non-research purposes. One example is the Swedish Twin registry , which covers Swedish twins born 1886 and onwards, and contains data on more than 200 000 individuals.[2]

A third characteristic of register data, related to the second, is that they seldom extend very far back in time. This is a consequence of the fact that many administrative tasks did not become effectively computerized until in the late 1980s. However, historical register data are becoming available to a steadily increasing extent. Data previously only available on paper, in files and binders, are being transferred to electronic media, enabling analyses covering long periods of time. One example of a register data base with long historical records is the Swedish register on tertiary students. Information about enrolled students goes back to the school year 1977/1978 and data on individuals with completed degrees to 1962/1963. Another example from Sweden is the Child welfare intervention register dating back to 1968.[3]

3 Ways to make use of register data

Essentially, there are three different ways to make use of register data:

As they come. In this case, typically, information from many different registers is combined. An example is provided in Section 7.

To enrich already collected data. In general, the already collected data will be some kind of survey data. For example, the researcher may have collected data on teachers and students that have been involved in a pedagogical support program. To evaluate the effects of the support program, the survey data need to be complemented with background data and data on comparable teachers and students that have not been affected by the program. The matching of the survey data with the register data requires unique identifiers for the teachers, students, and schools covered by the survey data. For an illustration, Section 8.

In connection with a data collection to be conducted. The aim of this way to use register data is to reduce the respondents' 'informational burden', by only asking them about facts and data which cannot be obtained from other sources. For instance, in a

[2] See, further *http://ki.se/en/research/the-swedish-twin-registry* , accessed July 2016. For analyses of the return to education based on the twin registry, see Isacsson (1999, 2004).

[3] For a study exploiting this register to analyze the relation between performance in primary school and psychosocial problems in young adulthood among individuals that have been placed in foster care, see Berlin et al. (2011).

study of, say, student attitudes, the students need not to be asked about their family backgrounds in terms of their parents' educations, occupations, and earnings, as information on these variables are readily available in register data bases. This use of register data not only has the advantage of making surveys shorter and, thus, cheaper; it also reduces non-response rates and, accordingly, increases the precision in the conclusions that can be drawn from the collected information.

4 Register data of interest – and information lacking

One way to describe the possibilities and limits of register data is to use different types of data classifications (taxonomies), showing categories/classes for which there are and are not register data, respectively. That will be the main approach pursued here. However, to complement this display of register data *availability* with information about the *interfaces* between different types of data two simple schematic figures will be provided, too.

Following Cedefop (2014), we start by considering the following three broad categories of education and learning:

> *Formal education and learning*:
> Learning that occurs in an organized and structured environment (such as in an education or training institution or on the job) and is explicitly designated as learning (in terms of objectives, time or resources). Formal learning is intentional from the learner's point of view. It typically leads to certification.
>
> *Non-formal education and learning*:
> Learning which is embedded in planned activities not explicitly designated as learning (in terms of learning objectives, learning time or learning support), but which contain an important learning element. Non-formal learning is intentional from the learner's point of view. It typically does not lead to certification.
>
> *Informal learning*:
> Learning resulting from daily activities related to work, family or leisure. It is not organized or structured in terms of objectives, time or learning support. Informal learning is in most cases unintentional from the learner's perspective. Also referred to as experiential or incidental/random learning.

Table 1 shows, in broad terms, to what extent there are register data on each of these three categories. 'Broad terms' refers both to the nature of the entries in the table, like 'tertiary education' and 'on-the-job training', and to the assigned associated availability

of register data (extensive/limited/non-existent); with respect to the latter there are also differences across the Nordic countries.

It can be seen from **Table 1** that only formal education and training are represented in register data in the Nordic coumtries. It should be noted, however, that information on non-formal education and training is collected in two large surveys, the Adult Education Survey (AES)[4] and the Programme for International Assessment of Adult Competencies (PIAAC)[5], both of which can be matched with register data.

[4] The AES is part of the EU Statistics on lifelong learning and covers people in the age range 25-64, of which representative samples of individuals are randomly selected for interview, in each country. The first and second waves of the AES were conducted in 2007 and 2011. The Nordic countries Denmark, Finland, Norway and Sweden participated on both of these waves. In 2011 the samples for these countries varied between 5 000 and 6 000 persons and the response rates were in the range 55–65%. The third wave is taking place in 2016. See further: *http://ec.europa.eu/eurostat/web/microdata/adult-education-survey* , accessed July 2016.

[5] The PIAAC is conducted by the OECD and targets individuals aged 16-65. Its primary purpose is to assess skills in literacy, numeracy and problem-solving by means of information and communication technology but extensive information is also collected on education and training. The first wave of PIAAC was conducted in 2011–2012 (23 countries) and 2014–2015 (nine countries). The Nordic countries Denmark, Finland, Norway and Sweden participated in 2011–2012. The country samples were collected so as to be representative of the adult population. In the four Nordic countries the number of respondents varied between approximately 4 500 (for Sweden) and about 7 300 (for Denmark). For detailed information, see OECD (2013) and OECD (2016).

Table 1: Forms of education and learning by representation in register data im the Nordic countries

	Type of education and learning		
Register data	*Formal* [1]	*Non-formal*	*Informal*
Extensive	Compulsory school Upper secondary school Tertiary education Labor market training Language courses for immigrants		
Limited	Pre-school Folk high school – long courses		
Non-existent		On-the-job training Folk high school – short courses Adult educational associations Study circles Seminars Private lessons	Self studies Every-day activities intended to promote learning

[1] No distinction is made here with respect to youth and adult education. Thus, e.g., compulsory school and upper secondary school include adult education to obtain compulsory school or upper secondary school qualifications.

Another useful taxonomy relates to the "observational units" as defined in register data. At the lowest levels of aggregation, the observational units of interest in the present context are individuals and education and training sites.

With respect to individuals, the relevant categories here are students/learners, teachers/instructors, managers/leaders, and support staff, cf. **Table 2**. The types of data listed in the table are not meant be exhaustive; they are merely providing examples of data that can be of interest for empirical educational science studies.

Table 2: Types of register data, by categories of individuals, in the Nordic countries

	Category of individuals			
Type of data	*Students and learners*	*Teachers and instructors*	*Management and leaders*	*Support staff*
Demographics	Sex, age, sibling and parent info	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*
Family status	Single, cohabiting, married, divorced, widow/widower	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*
Immigrant status	Country of birth, mother tongue, host country language training	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*
Geographical info	Place of residence, location of school/ learning premises	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*
Education/training	Level + field-of-study for highest & latest education attained, ongoing formal education/ training, grades, results on national tests	Cf. *Students and learners* + teacher certificate: general and/or by subject	Cf. *Teachers and instructors*	Cf. *Students and learners* or *Teachers and instructors,* depending on applicability
Labor market info	Work experience, tenure, unemployment, occupation, wages, earnings	Cf. *Students and learners* + Experience and tenure as teacher, subjects taught	Cf. *Teachers and instructors*	Cf. *Students and learners* or *Teachers and instructors,* depending on applicability
Income	Earnings, capital income, transfers: unemployment & sickness benefits	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*
Health information	Sickness leave, in-patient care, disabilities, diagnoses	Cf. *Students and learners*	Cf. *Students and learners*	Cf. *Students and learners*

It should be noted that the frequencies at which data are recorded differ across the various types of information displayed in **Table 2**. Some pieces of information are only recorded once, like country of birth and biological parents, while others are up-dated on a daily basis, like data on unemployment and sickness benefits. In between these extremes are yearly data on, e.g., education and place of residence, and quarterly data on, i.a., employment and earnings.

The registers do not contain more qualitatively oriented information like, e.g., attitudes, ambitions etc. This does not necessarily mean that such aspects cannot be accounted for, however. As register data are made up of repeated observations on the same individuals, i.e. *panel data*, aspects like attitudes and ambitions can be controlled for, as long as these aspects are constant over the time period studied. For example, consider individual skills. These may be dependent on education and training, attitudes towards schooling, and learning ambitions.[6] If the individuals' education and training change during the study period while their attitudes and ambitions remain fixed, the impacts of education and training on skills, *controlling for attitudes and ambitions*, can be estimated by relating changes in skills to changes in education and training. Thus, there is no need for explicit measures of variables like attitudes and ambitions when these do not vary, implying that the corresponding changes are zero; see further Hsiao (1986). But, of course, in some cases, changing students' attitudes and ambitions may be (part of) the objectives of an educational reform. In such cases the register data have to be complemented by survey data on these aspects.

Another kind of relevant information that is not covered by register data concerns school and teaching support programs that are either local or regional in nature. At the local government level, and even in individual schools, it is quite common to find small-scale initiatives to improve teaching methods or to try out alternative methods of organization like, e.g., team-work approaches. These efforts will not be recorded in register data (although they are often locally documented). Even quite large and extensive teaching support schemes may go unnoticed in register data, if they have been initiated outside the nationally financed system of education and training. In Section 8 an example of such a scheme is considered, the Swedish Natural sciences and Technology for All (NTA) Program.

[6] For simplicity, important aspects like family background are disregarded here.

Regarding workplaces, the natural entities are various kinds of ’learning sites’. In the context of formal education and training these are pre-schools, schools, colleges and universities, and training facilities for, e.g., labor market training and immigrant language courses. For non-formal and informal training, learning sites may sometimes be identified – on-the-job training is a case in point, where the site can be the individual’s workplace – while in other cases information about the learning site may be lacking – in the case of study circles and self-studies, for instance.

Table 3 provides a categorization of the different types of data that are available for the relevant workplaces, illustrating these categories by means of examples.

Table 3: Types of data for education and training premises

Main category	Sub-categories; examples
Type of education/training provided	Pre-school, compulsory school, university, on-the-job training
Student intake	In total, by grades, by programs, by fields-of-study
Students enrolled	In total, by grades, by programs, by fields-of-study
Student examinations	In total, by grades, by programs, by fields-of-study
Employees	Teachers, management, support staff
Expenditures	Salaries, rents, materials

As a rule, the information in **Table 3** is recorded annually.

While Tables 1–3 provide an overview of the different kinds of information that are contained in registers they do not provide guidance about how these types of data can be linked to one another. This aspect is considered in **Figure 1** and **Figure 2**.

Figure 1 shows that information about the different categories of individuals in **Table 2** can be connected with information about the education and training sites. For example, for a given university it is possible to extract information on its students and various categories of staff. And vice versa, students and teachers/school leaders/support staff can be linked to the education or training institution where they are active. However, there are no horizontal links in the figure, from one category of individuals to

another. This means that the different categories of individuals can only be indirectly connected, through education and training facilities. In particular, students and learners cannot be directly connected with teachers and instructors. Accordingly, register data are not informative about which teacher(s) the student/learner has been taught by. For educational science purposes, this is the main drawback with register data; indeed, innumerable studies have concluded that the teacher is *the* most important determinant of the student's learning outcomes, see, e.g., Gustafsson (2003) and Hattie (2009). Thus, if combining survey data and register data is an option, presumably the most valuable survey data that can be collected is information about which teacher(s) that taught a given set of students. An example of such a combination of survey and register data is provided below.

Figure 1: Links that can be established between observational units in register data

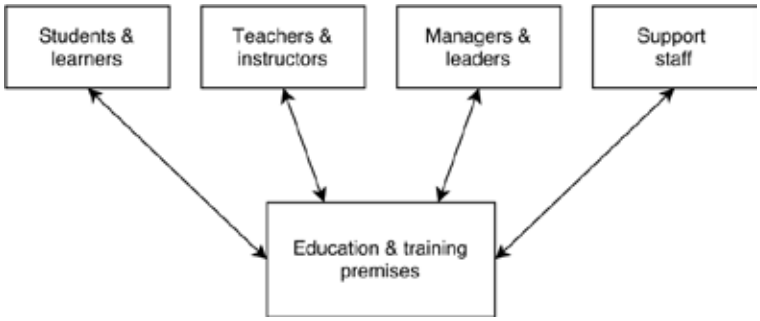
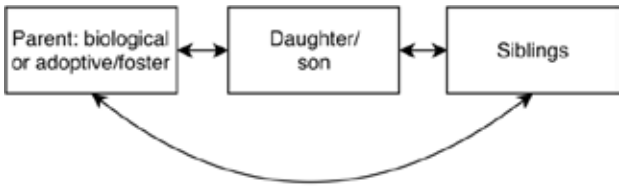


Figure 2 describes the intergenerational links available in register data. It should be noted that while the figure only illustrates the link between two consecutive generations, register data may contain links between three or even four generations.[7]

The importance of the intergenerational links is that they make it possible to account for different aspects on 'nature and nurture'. An obvious example is the well recognized importance of parental characteristics in studies of, e.g., children's learning behavior. Another example: to evaluate the influences of various kinds, or amounts, of schooling, one can study differences between siblings that have been brought up together but have not had the same educational experience, thereby controlling for home environment and genetic factors.[8] An alternative way to eliminate genetic factors is to study adoptees or foster children.

[7] In general, the number of generations is determined by the upper age limit that the reisters are subject to.
[8] To fully control for genetic factors in this way one has to consider a very specific group of siblings, namely identical twins.

Figure 2: Intergenerational links in register data



5 How to gain access to register information

The first thing to note about register data is that, in general, they are available for research purposes only. To verify that the data are going to be used within a research project, a request for a register data set has to be accompanied by a research plan, containing research question(s) and a description of the data and methods required to answer the question(s).

Given the research plan, the researcher first has to decide on whether the project needs to undergo a review by an *Ethical Committee*. The purpose of the ethical review is to weigh the expected social benefits and knowledge gains generated by the project against the risks for, and costs associated with, violations of personal integrity

For projects employing register data only, an ethical review may not be necessary. However, if the project's data involve health information an ethical review is *always* required. This derives from the fact that in medicine, in contrast to the social sciences, all research projects have to be approved by an Ethical committee before they can be initiated, cf. Ludvigsson et al. (2015). Since, with respect to ethical considerations, the same rules should apply in the social sciences as in medicine, this is tantamount to saying that an ethical review is required if the data include health information (irrespective of whether this information has been extracted from register databases or has been obtained in some other way).

When the project dataset is constructed by register data *and* survey data an ethical review is almost always required. This is because of the requirement of *informed consent* that arises in connection with survey data. The ethical review is to ascertain that the persons targeted by the survey are appropriately informed about i) the research project, in terms of its aims, methods, data, and publication of results, ii) why and how persons have been selected to participate in the survey, iii) the project's handling of data

and presentation of results so as to secure personal integrity, iv) that participation is voluntary, and v) how to accept or decline participation in the project.[9]

A distinction is made between *active* and *passive* informed consent, respectively. Active consent means that the prospective participant explicitly, in writing, accepts or declines participation. Passive consent, on the other hand, means that unless the persons explicitly (in writing) declines participation (s)he implicitly agrees to participate. In deciding between active and passive consent the Ethical Committee considers, i.a., that if the study involves a very large number of individuals active consent may by infeasible, and that active consent may sharply reduce response rates and thereby the precision in the project's results.

If the survey is to be conducted as part of the project it may seem that by responding to the survey the person agrees to participate in the project, otherwise (s)he declines participation. However, the important aspect here is that the survey data are combined with register data; cf. the discussion above under the heading *Ways to make use of register data*. If part of the information about the participating persons is obtained from register data, informed consent about the register data part of the data collection is necessary. The same applies if the project employs a survey that has been conducted earlier, prior to the project, and complements the survey data with register data. Accordingly, survey information and register data cannot simply be merged; the researcher has to plan ahead, inform the respondents about the intention to combine the two types of data and obtain their consent for doing so.

In addition to being prompted by the issue of informed consent, an ethical review may be required for other reasons, too. Examples relevant in the educational science context are projects that concern children and/or involve video recordings.[10]

When the project has been approved by the Ethical Committee (given that an ethical review is required), the next step is to approach the organization(s) administering the data of interest. In general, this will involve contacting the national statistical agency[11]

[9] Survey data may also include information classified as 'sensitive' in the Danish, Icelandic, Norwegian, and Swedish Personal Data Acts. In addition to health information, data on political, religious, and sexual disposition are classified as sensitive. Inclusion of these types of data, by itself, calls for an ethical review.

[10] When in doubt about whether an ethical review is required it is recommended to contact the country's regional or central ethical committees. The web-addresses of the central committees are: Denmark: *www.cvk.sum.dk*, Finland: *www.tenk.fi*, Iceland: *www.vsn.is*, Norway: *www.etikkom.no*, and Sweden: *www.epn.se*, accessed July 2016.

[11] Statistics Denmark, Statistics Finland, Statistics Iceland, Statistics Norway and Statistics Sweden.

but it may also involve other agencies. For example, health-related data are, for the most part, not administered by the national statistical agencies.[12]

Before providing the researchers access to the requested dataset, the data administering organization(s) make an assessment about the harm that the project potentially may do to the informants if their personal integrity is violated, in terms of, e.g., costs, (bad) reputation, disclosure of organizational or personal confidential information, etc. Personal integrity cannot be violated by disclosure of specific pieces of individual information because the data are pseudonymized before they are released to the researchers. This means that the unique personal and workplace identifiers have been replaced by serial numbers; the information needed to recover the original identifiers – the key – is kept by the national statistical agency. However, in rich datasets it may still be possible to locate specific individuals by means of backward identification, i.e. by combining many pieces of detailed information.[13]

The assessments made by the data administering organizations are based on laws and regulations that partially differ from those underlying the ethical reviews. Accordingly, the fact that a project's dataset has been considered acceptable by the Ethical Committee does not guarantee that it will approved by the agencies supplying the data. As a rule, the decision made by the data administering organization(s) is similar to the decision of the Ethical Committee but sometimes somewhat more cautious and, thus, somewhat more restrictive.[14] It may be that part of the register information will be provided in more aggregated form than originally requested, in order to make backward identification more difficult. For instance, individuals' country of birth may be replaced by region of birth.

Sometimes it may be possible to extract the data from register databases that have been constructed by research institutes. Two examples in the Nordic countries are the Danish National Centre for Social Research (SFI) and the Institute for Evaluation of Labour Market and Education Policy (IFAU), in Sweden. When feasible, data access through these institutions is cheaper and faster than through the national statistical agencies. However, in general, the underlying research project has to (partly) involve

[12] Information about which institutions to contact can be obtained from national statistical agencies.
[13] Backward identification is illegal and researchers handling pseudonymized data are often required to acknowledge, in writing, that they are aware of this being the case.
[14] Presumably, this is due to the fact that the data administering organizations, unlike the ethical review committees, only consider the risks of violating personal integrity and not the project's social benefits.

researchers employed at these institutes. Alternatively, in the case of IFAU, the project does not have to include any IFAU researchers, provided that it is supported by an IFAU research grant; these grants can be applied for also by researchers in the other Nordic countries. However, in terms of the discussion under the heading *Ways to make use of register data*, the utilization of data obtained through this channel is limited to the 'as they come' option, because the data are pseudonymized.

To ascertain safe handling of the data and prevent them from being accessed by unauthorized persons, the researchers are usually granted *remote access* only.[15] This means that, physically, the data never leave the data administering organization and that access to them is safe-guarded by means of a log-on procedure requiring the user to provide a specific password. Moreover, there are time limits to accessibility, as well. After a pre-specified period of time the remote access channel will be closed, and the key will be destroyed, unless the terms of the original agreement are re-negotiated.

Finally, a remark is in place regarding the fact that the discussion above has concerned access to *intra-national* register data only. It is not unlikely that register data for all of the Nordic countries will become available in the near future. Indeed, a successful first attempt to create a database containing register data for Denmark, Finland, Norway, and Sweden has already been conducted. The core of the database is made up of data from the PIAAC survey to which register data for 2011 have been added, containing, i.a., geographical information, labor market data, and take-up of social welfare benefits. For information about how the database can be accessed, see Rosdahl (2015).

6 Combining quantitative and qualitative research

In educational science, as in many other disciplines, quantitative and qualitative methods are seldom combined. It seems, however, that empirical educational science research can benefit from merging qualitative and quantitative approaches.[16] This will be illustrated by means of a fictional example.

[15] At the time of writing this article remote access was not in use in Norway; instead the researchers were provided with copies of the data that were stored locally.

[16] Of course, educational science is not the only discipline foregoing the opportunities provided by multi-methodological approaches.

Assume that female students are believed to be under-represented in upper secondary school science tracks. Moreover, it is conceived that the female representation can be increased by appropriate teaching methods and student counseling. How to test this conjecture? Stereotypically, the research strategies of a qualitatively and a quantitatively oriented researcher, respectively, could be described as follows.

To confirm that female students are under-represented, the qualitative researcher checks a yearbook of educational statistics, which shows that girls make up less than 50 percent of the science students. Next, the researcher asks the teachers' union for examples of upper secondary schools that have been successful in attracting girls to science tracks. Among the schools suggested, two are chosen, both within commuting distance but located in poor and well off neighborhoods, respectively. At each school, 15 randomly selected female science students are interviewed about why they have chosen a science track. The girls' science teachers and study counselors in the last year of compulsory school (identified by the girls) are also interviewed.

The interviews provide information about teaching methods and counseling advice that appear to increase the girls' interest in science and make them choose science tracks, in two quite different socio-economic contexts. Two objections might be raised, however. First, the measure used of under-representation is rather simplistic. Second, the girls interviewed may be a selective group. Specifically, it may be that they have properties – upbringing, interests, ambitions and study skills – such that they would have chosen science tracks in upper secondary school irrespective of which compulsory school they had happened to attend. If so, the science teachers and study counselors in the compulsory schools that the girls attended may not be representative of teachers and counselors that are successful in making girls choosing science tracks.

The quantitative researcher starts by compiling register data, for all students attending the last year in compulsory school, on sex, family background, school, and scholastic achievements (grades and results on national tests) and choice of track in upper secondary school. Regression analysis is then used to model the choice between science tracks versus other tracks.[17] The analysis provides a measure of girls' under-

[17] For example, a logistic regression can be applied; see, e.g., Hosmer & Lemeshow (2000). In this regression equation the dependent variable is binary: 1 if the student chose a science track in upper secondary school and 0 otherwise. The explanatory variables, whose impacts on the dependent variable are to be estimated, are binary variables indicating sex, school and the products of these two, plus non-binary variables representing family background and scholastic achievements.

representation in science tracks that is reasonable in the sense that it controls for parental background and scholastic achievements.[18] Moreover, it yields information about which compulsory schools that matter for students' choice of tracks in upper secondary schools and, in particular, if they increase or decrease the likelihood of girls choosing a science track.[19] The problem with the analysis is that it cannot say anything about why.[20]

However, the two approaches can be combined. Specifically, the results of quantitative analysis can be used to determine which schools to be studied more closely, i.e. qualitatively, namely the schools that have been found to matter – positively or negatively – for girls' choice of science tracks. The qualitative analysis of these schools will provide answers to why those schools matter.

It should be noted that this combined use of quantitative and qualitative methods rests on them being applied in a specific order: the quantitative analysis should precede the qualitative. This feature is not special to this example but holds in general. The idea is to employ the quantitative analysis as instrument for selecting the case studies which are most likely to provide insights about the mechanisms underlying the patterns observed in the quantitative data.

7 An example of research requiring register data

There are quite a few research problems that can only be credibly examined by means register data, simply because other sources cannot provide sufficient numbers of observations. The example considered here concerns school starting age.

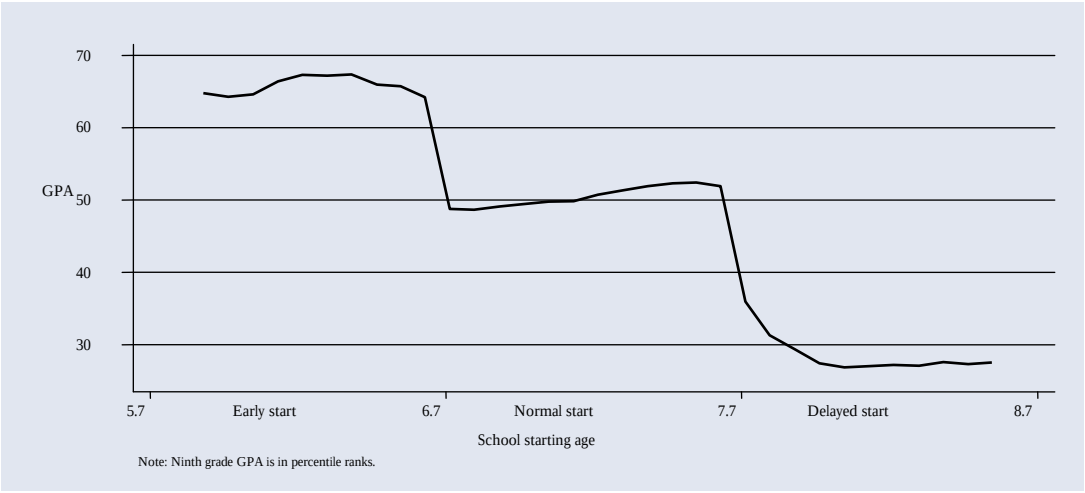
In an international perspective, children in the Nordic region start school quite late. The (normal) starting age used to be 7 in Denmark, Finland, Norway, and Sweden. However, since 1997 the starting age is 6 in Norway, see, e.g., Mellander & Fremming Anderssen (2015) This prompts the question: Is it better to start school earlier? Using

[18] Measured by the estimated parameter for the binary sex variable, coded 1 for girls and 0 for boys.
[19] By means of the estimated parameters for the sex × school variables.
[20] Another problem is that the students implicitly are assumed to be *assigned* to compulsory schools, in accordance with, e.g., a neighborhood criterion. If, instead, the students themselves, or their parents, can choose which school to attend a selection problem arises, similar to one discussed in connection with the discussion of the qualitative approach. In that case, the analysis outlined needs to be preceded by an analysis of school choice. However, this complication adds nothing to the argument here and is, therefore, disregarded.

Swedish register data, Fredriksson & Öckert (2006, 2014) provide an answer to this question.[21]

Across the 1975-1983 birth cohorts the actual school starting age in Sweden varied between approximately 5.8 and 8.6 years, cf. **Figure 3**.

Figure 3: School starting age and grade point average (GPA) at the end of compulsory school in Sweden, for the 1975-1983 birth cohorts



Source: Fredriksson and Öckert (2006).

In **Figure 3**, the relation between school starting and grade point average, GPA, is negative – children starting school earlier have higher GPAs than children starting later. At first sight, this would seem to indicate that it is better to start school earlier. However, it does not take too much thought to realize that the school starting age is not random – 'bright' kids tend to be over-represented among the children starting early, while less talented children tend to be delayed. Moreover, it is important to separate differences in age from differences in school starting age. Specifically, one would like to compare children that differ as little as possible with respect to chronological age but, nevertheless, differ with respect to school starting age. In principle, this is possible, because children born right at the end of year t will start school in the year $t + 7$ while children born in the beginning of year $t + 1$ will start school one year later. In practice, though, such a comparison is very difficult to conduct if data can only be collected by means of surveys – it would be very difficult and costly to find sufficient numbers of children born in, say, December and January.

[21] For similar analyses, relating to Norway and England, respectively, see Black et al. (2011) and Crawford et al. (2010).

With access to register data, this is not a problem. According to **Table 4**, register data for the 1975-1983 birth cohorts contain over 66 000 children with normal school start that were born in January.[22] For December the corresponding number is over 51 000.

Table 4: Month of birth and timing of school start (percentage points) for the Swedish 1975-1983 birth cohorts

	January	February	March	April	May	June
Earlier school start	2.58	1.28	0.81	0.61	0.40	0.32
Normal school start	96.64	97.87	98.22	98.31	98.24	98.06
Delayed school start	0.79	0.86	0.97	1.08	1.36	1.63
n	66,361	65,816	77,092	75,932	72,516	66,912
	July	August	September	October	November	December
Earlier school start	0.25	0.15	0.14	0.10	0.06	0.02
Normal school start	97.94	97.72	97.38	96.88	95.41	91.14
Delayed school start	1.81	2.13	2.49	3.02	4.52	8.84
n	67,394	65,036	64,987	61,507	56,215	56,560

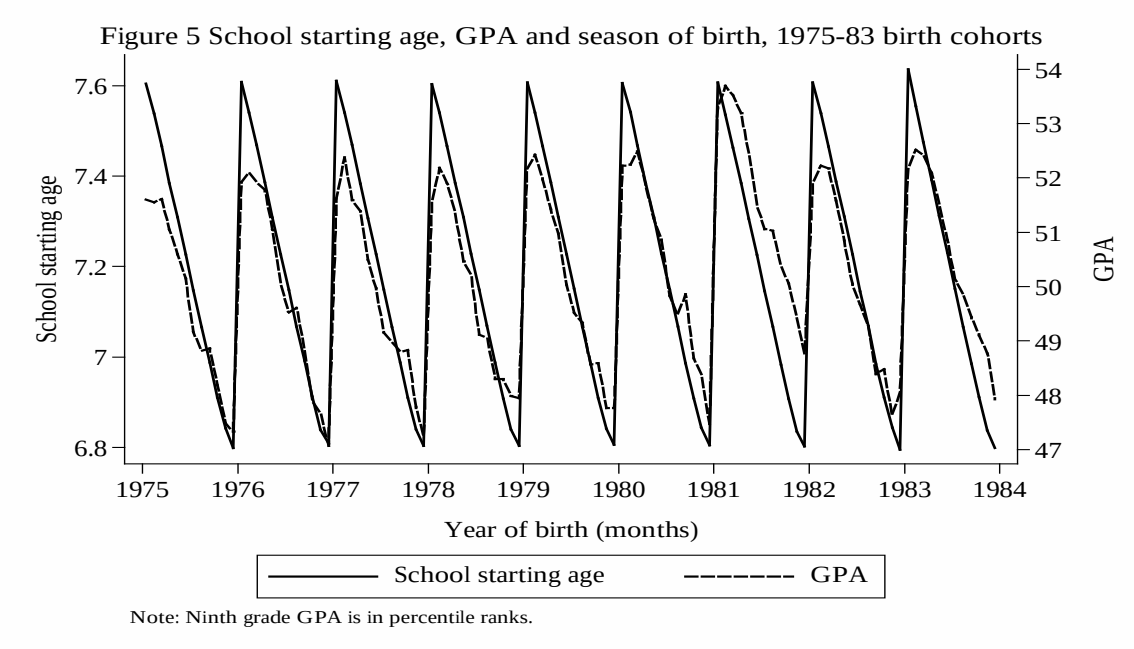
Note: Normal school starting age is between 6.8 and 7.7 years depending on month of birth.

Source: Fredriksson and Öckert (2006).

Now, consider **Figure 4**, showing the relation between the normal school starting age and GPA. It is quite clear that when we compare the children born at the end of the year with children born at the beginning of the year, the relation between school starting age and GPA becomes positive – it seems better to start school *late*r, not earlier. This conjecture is verified by Fredriksson and Öckert (2006), who also show positive effects of later school start on educational attainment and on the probability of attaining a college degree.

[22] Normal school start means school start in year t +7 for children born in year t. According to Table 4, the number of children with normal school start that were born in January is given by 0.9964 × 66 361.

Figure 4: School starting age, GPA at the end of compulsory school, and season of birth, for the Swedish 1975-1983 birth cohorts



Source: Fredriksson and Öckert (2006).

8 An example of research using merged survey and register data

The section *Register data of interest – and information lacking* above points to two reasons for merging survey and register data. First, there are certain types of information that are not covered by register data. Second, in register data links between teachers and students are missing. This section describes an ongoing empirical study which relates to both of these reasons. The object of the study is to evaluate the effects of the teacher support program Naturvetenskap och Teknik för Alla[23], NTA, on scholastic achievements in natural sciences, measured in terms of grades and results on national standardized test in the 6th school year.

NTA is an inquiry-based, constructivist, program, built around a number of experiment tool kits (‘boxes’). In 2016, NTA was used in 128 of Sweden’s 290 towns and municipalities, including Sweden’s two largest towns, making it, by far, the biggest teaching support program in Sweden.

The NTA is not a part of the Swedish regular system of education. The program is administered by the organization NTA School Development. The only way to obtain

[23] In English: Natural sciences and Technology for All.

information about NTA activities and NTA participants is through NTA School Development; there are no records in register data.

In an earlier study of the effects of NTA, based on a random sample of 15 000 students in the 9th year of compulsory school, statistically significant positive effects of NTA were found on the results in national tests in Physics, but not in Biology and Chemistry. No statistically significant effects at all were found for course grades, for any of the three subjects.[24]

That only one significant effect of NTA was found may, of course, reflect the fact that the NTA is not a very efficient teacher support program. However, there may be other explanations, too, which relate to the data. First, the sample was, in practice, quite small; among the 15 000 students only 1 000 had participated in NTA and these were compared to the 1 000 of the 14 000 non-participants that were most similar to the NTA participants – 'synthetic twins'.[25] Accordingly, the effect evaluation was essentially based on 2 000 students only. Second, there was, in general, a considerable time lag between the participation in NTA and the effect evaluation. The NTA program is primarily employed in school years 4-6 while the effect evaluation was conducted in the 9th school year. Thus, the effects, if any, could (partly) have faded away during the three year period between the years when most of the students participated in the NTA and the year for which the effects were evaluated. Finally, no account could be taken of the fact that there were differences across the NTA-students with respect to the extent to which they had participated in the program and which NTA teacher(s) they had had.

The ongoing study addresses all of these issues. During the years 2011-2014, NTA School Development has collected data on all schools that make use of the NTA program, the participating students and their teachers, by semester, for the school years 4-6. The data comprises about 40 000 students. Information is also included about the particular experiment boxes used, making it possible to characterize them in terms of the natural science subjects Biology, Chemistry, and Physics. The NTA effects will be evaluated for the school years 2012/2013 and 2013/2014, in terms of the results on national tests and grades in the three science subjects in the 6th school year, i.e. right after the period during which the students have participated in the NTA program.

[24] On this study, see Mellander & Svärdh (forthcoming, 2015).
[25] These 1 000 non-participants were identified by means of Propensity Score Matching, a method based on regression analysis, cf. Guo & Fraser (2010).

With informed consent from the NTA teachers, the NTA students and their parents the survey data collected by the NTA School Development program is matched with register data, by means of the unique personal identification numbers of the NTA students and teachers, and the unique workplace codes of the NTA schools. This yields information about the family background and scholastic achievements in the 3rd school year (national test results in Swedish and Math) and in the 6th school year (national test results and grades in the natural science subjects, and test results and grades for several other subjects) for the NTA students. The national test results and grades in natural science subjects in the 6th year are outcome variables while the 3rd year test results control for pre-treatment scholastic aptitude and the test results and grades in non-science subjects in the 6th year control for post-treatment scholastic aptitude. Moreover, the register data contain detailed information about the education and work experiences of the NTA teachers, the schools that the NTA students attended and the municipalities and regions where the schools are located.

All of the information about the potential comparison group – students which did not participate in NTA during the period 2011-2014 but attended 6th grade in either of the school years 2012/2013 or 2013/2014 – is extracted from register data. Moreover, the registers also provide information about the schools that these students attended and the municipalities where the schools were located.

The assessment of the effects of the NTA program proceeds in two steps. First, non-NTA students acting as synthetic twins to the NTA participants are selected, among the students in the potential comparison group. In the second step the effects of the NTA program are determined by comparing the test result and grades in the natural science subjects in the 6th school year between NTA and non-NTA students. The subject effects will be allowed to vary with the number of semesters that the NTA students have participated in the program, the kind of NTA boxes used, and the characteristics of the NTA teacher.

9 Evaluations of causal effects of educational interventions

An important application of register data is evaluations of causal effects of educational interventions, i.e. the establishment of co-variation that goes beyond correlations and identifies a direction of impact. This issue has already been touched upon, in the two

preceding sections. In the first of these, the causal effects of school starting age on scholastic achievements were estimated, whereas the other concerned the effects of the Swedish NTA program, again on scholastic achievements. As these two examples show, causal effects can be estimated by means of register data only (the first example) or by means of register data in conjunction with survey data (the second example).

The essential feature of a causal effect evaluation is that it identifies a *counterfactual* treatment or event that describes what would have happened, had the intervention not taken place. Valid counterfactuals can only be defined if the implemention of intervention satisfies certain conditions. Three different conditions can be distinguished: randomized experiments[26], natural experiments and quasi-experiments.

In randomized experiments, the 'treatment', i.e., the intervention, will not comprise all individuals that might benefit from it, but only a randomly selected subset of them. The individuals in the other, not selected, subset constitute the counterfactuals; they act as a control, or comparison, group. The effect of the intervention is simply determined by comparing the average outcomes in the treatment group with the average outcomes in the control group.

Randomized experiments are very rare in education and training. One example is an intervention in Danish pre-schools, analyzed by Jensen et al. (2013). A teacher support program comprising different aspects on pre-school pedagogics was randomly assigned to comparable pre-schools in two Danish municipalities. Both register data and survey data were employed in the effect evaluation. The results showed that the treated children exhibited fewer emotional problems, were less hyperactive and more attentive than the children in the control group. In an evaluation of the returns to college in Sweden, Öckert (2010) provides another example of a randomized experiment: at the margin, some students were granted entry by means of lotteries.

Natural experiments are events that, by change, have taken place in such a way that they can be regarded as experiments, although no randomization has occurred. The study about the effects of school starting can be viewed as a natural experiment, exploiting the fact that the timing of births immediately before and after December 31st, respectively, can be considered as random.

[26] Also denoted randomized control trials, or RCTs.

In the Nordic countries, the most well-known examples of natural experiments concern the introduction of compulsory school. In Sweden, Norway and Finland the corresponding policy reforms were implemented progressively, with respect to both time and space. As a result, during the process, similar municipalities could be observed that had and had not, respectively, implemented the new system. The effects of the reform were estimated by comparing the achievements of students of the same age that, in parallel, attended schools were the reform was implemented early and later, respectively, cf. Meghir and Palme (2005), Pekkarinen et al. (2009) and Aakvik et al. (2010). One effect was a small increase in the average years of schooling. Life-time earnings were also affected – for the better for children with low-educated parents and for worse for children with well-educated parents.

In contrast to natural experiments, quasi-experiments refer to situations where there is no random distribution across the treatment group and the control group. The planned effect evaluation of the NTA program will be carried out in the form of a quasi-experiment. Specifically, students which did not participate in the program but are very similar to the participating students in all other relevant respects will be selected by means of a statistical matching method (Guo and Fraser, 2010).

Another example of a quasi-experiment is Hall's (2012) evaluation of the prolongation of the Swedish upper secondary vocational education from two to three years. Like the compulsory school reform, this reform was implemented gradually making it possible to compare students attending three-year programs with similar students attending the corresponding two-year programs. To account for the non-random selection into the two types of program Hall (op. cit.) employed a so called instrumental variable (IV) method; cf. Cameron and Trivedi (2005). The IV method substitutes the students' actual program choices by an indicator – an instrument – that is correlated with the actual choices made but not with either the outcomes of the programs or the student's characteristics. Hall's (op. cit.) instrument was the availability of the three-year programs in the municipalities considered, measured as the proportion of three-year programs among all upper secondary vocational programs. The effect evaluation was based on register data only and showed that the prolongation of the vocational programs increased the likelihood of students dropping out of upper secondary and had no effect on the transitions to higher education.

10 Concluding discussion

To give a full account of the wealth of register data in the Nordic countries is an insurmountable task. Fortunately, it not desirable to do so, because users are interested in limited subsets of the register universe. Nevertheless, to be able to use register data efficiently, it is necessary to understand the basic principles upon which they are built. In this article, this dual need for overview and detail has been resolved by describing register data both in very general terms and in rather specific terms, the latter focusing on informational content from an educational science perspective. Moreover, a quite extensive discussion has been devoted to the issue of getting access to register data, because this is where most first time users stumble.

Oftentimes, the best way to promote a new instrument – like register data for (most) educational scientists – is to show, by means of concrete examples, what it can do. To this end, practical applications have been discussed, each of which illustrate advantages of using register data, as opposed to relying on interview and survey information only.

The article's main conclusions can be summarized as follows:

First, register data contain lots of interesting information for educational scientists. In addition, the fact that register data are *panel data* – repeated observation on the same individuals and organizations – makes it possible, in analyses, also to control for aspects on which the registers are not informative, provided that these aspects are unchanged over the period under study. Register data also contain intergenerational links, making it possible, e.g., to study the influence of family background on learning and, furthermore, to separate genetic and environmental factors.

Second, register have one important shortcoming: they cannot tell about direct connections between students/learners and teachers/instructors. Only indirect links, via the school, university, or learning site are available. However, this problem can be overcome by matching register data with survey data containing information about which students that have been taught by which teachers.

Third, simply by containing very large numbers of observations, register data enable analyses which are very hard – or impossible – to address by means of interview or survey data.

Fourth, combinations of register based statistical analyses and in-depth case studies will provide benefits for both qualitatively and quantitatively oriented researchers.

Fifth, register data are instrumental in evaluations of causal effects of educational interventions.

Another relevant issue concerns the costs of register data, compared to interview or survey data. In general, register data are cheaper, due to the fact that their compilation involves much fewer work hours than the collection and structuring of interview and survey data – hours which the researcher employing register data can devote to other tasks. Mostly, this difference more than outweighs the costs involved when information is extracted from registers.[27]

Furthermore, as noted by Ludvigsson et al. (2015), data for different registers are compiled independently, which minimizes bias in data collection. And in studies with long time follow-up, outcome information from registers is more reliable than survey information which may uncertain due to imprecise personal recall or attrition.

[27] Of course, these costs vary both by the size of the data set and the nature of the information that it contains, and also by country. But to give an example, a rough estimate of the cost for the register data considered in the previous section (including the merging of those register data with the collected survey data) would be € 11 000 – 13 000.

References

Aakvik, A, Salvanes, K. G. & Vaage, K. (2010), Measurng heterogeneity in the returns to education using an education reform. *European Economic Review, 54*(4), 483500.

Berlin, M., Vinnerljung, B. & Hjern, A. (2011). School performance in primary school and psychosocial problems in young adulthood among care leavers from long term foster care. *Children and Youth Services Review, 33*(12), 2489-2497.

Black, S., Devereux, P. & Salvanes, K. (2011). Too young to leave the nest? The effects of school starting age. *Review of Economics and Statistics*, 93(2), 455-467.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics. Methods and Applications*, New York: Cambridge University Press.

Cedefop (2014). *Terminology of European education and training policy: a selection of 130 items, 2nd ed*. Luxembourg: Publications Office.

Crawford, C., Dearden, L. & Meghir, C. (2010). When You Are Born Matters: The Impact of Date of Birth on Educational Outcomes in England. IFS Working Paper, W10/06. London: Institute for Fiscal Studies (IFS).

Fredriksson, P. & Öckert, B. (2006). Is early learning really more productive? The effect of school starting age on school and labor market performance. IFAU Working Paper 2006:12. Uppsala: Institute for Evaluation of Labour Market and Education Policy (IFAU).

Fredriksson, P. & Öckert, B. (2014). Life-cycle effects of age at school start. *Economic Journal, Vol. 124*, 977–1004.

Guo, S. Y. & Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. New York: SAGE.

Gustafsson, J.–E. (2003). What do we know about effects of school resources on educational results? *Swedish Economic Policy Review, 10*, 77–110.

Hall, C. (2012). The Effects of Reducing Tracking in Upper Secondary School: Evidence from a Large-Scale Pilot Scheme. *Journal of Human Resources, 47*(1), 237269.

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Hsiao, C. (1986). *Analysis of panel data*. New York: Cambridge University Press.

Isacsson, G. (1999). Estimates of the return to schooling in Sweden from a large sample of twins. *Labour Economics, 6*(4), 471-489.

Isacsson, G. (2004). Estimating the return to educational levels using data on twins. *Journal of Applied Econometrics, 18*(3), 99-119.

Jensen, B., Holm, A. & Bremberg, S. (2013). Effectiveness of a Danish early pre-school program: A randomized trial, *International Journal of Educational Research, 62*, 115128.

Ludvigsson, J. F., Håberg, S. E., Knudsen, G. P., Lafolie, P., Sarkkola, C., von Kraemer, S., Weiderpass, E., & Nørgaard, M. (2015). Ethical aspects og registry-based research in the Nordic countries. *Clinical Epidemiology, 7*, 491–508.

Meghir, C. & Palme, M. (2005). Educational Reform, Ability, and Family Background. *American Economic Review, 95*(1), 414–424.

Mellander, E. & Fremming Anderssen, A. (2015). An overview of the characteristics of the Nordic region; Chapter 1 in *Adult Skills in the Nordic Region: Key Information-Processing Skills Among Adults in the Nordic Region*, TemaNord 2015:535, Nordic Council of Ministers, Copenhagen: Rosendahls-Shultz Grafisk.

Mellander, E. & Svärdh, J. (forthcoming). Tre lärdomar från en effektutvärdering av lärarstödsprogrammet NTA (Three lessons from an effect evaluation of the Swedish Science and Technology for Children program). *Nordic Studies in Science Education (NorDiNa).*

Mellander, E. & Svärdh, J. (2015). Inquiry-based learning put to test: long-term effects of the Swedish Science and Technology for Children program, IFAU Working Paper 2015:23. Uppsala: Institute for Evaluation of Labour Market and Education Policy (IFAU).

OECD (2013). *OECD Skills 2013: First results from the survey of adult skills*. Paris: OECD Publishing.

OECD (2016). *Skills Matter: Further Results from the Survey of Adult Skills, OECD Skills Studies*. Paris: OECD Publishing.

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics*, *93*(7-8), 965–973.

Rosdahl, A. (2015). PIAAC Nordic Database Guidelines_15 October 2015_DRAFT. Mimeo. Copenhagen: The Danish National Centre for Social Research (SFI).

Vetenskapsrådet (2015). Forskningens framtid! Ämnesöversikt 2014: Utbildnings-vetenskap. (The future of research! Overview of Educational science 2014). Stockholm: Vetenskapsrådets rapporter 2015. Stockholm: Vetenskapsrådet.

Öckert, B. (2010). What's the value of an acceptance letter? Using admissions data to estimate the return to college. *Economics of Education Review*, *29*, 504–516.