NSM 2022
NORDIC STATISTICAL MEETING

Matched Educational Data: Methods for matching and analysis

Alex Skøtt Nielsen, Statistics Denmark, axn@dst.dk

Jens Bjerre, Statistics Denmark, jbe@dst.dk

Abstract

Statistics Denmark is currently developing several new registries of 'Matched Educational Data' (MED), i.e., relational data that links students, teachers and activities together. These registries are of great interest to policy makers and researchers, since they allow for a more detailed understanding of teacher effects, peer effects and much more. This paper presents MED-products under development as well as two available products already in use by researchers.

A central purpose of the MED-registries is to provide data for the study of teacher effects. To this end, Statistics Denmark is currently using administrative school data to develop a registry of school activities covering all school subjects and non-formal activities (e.g. school trips) across the school year from 2020 and beyond. To capture teacher effects historically, Statistics Denmark employ two already available data sources 1) Statistics Denmark's registry of primary school students and 2) the Danish Ministry of Children and Education's data of primary school teachers formal teaching competency and planned teaching hours.

Another application of the MED-registries is the study of peer effects at the individual and group level. With the before mentioned administrative school data it is possible to study peer effects at the level of groups, subjects and even at the level of singular activities. To cover peer effects historically, Statistics Denmark has produced a 'Class-ID' linking groups of primary school students together across time (school years) and space (schools). The Class-ID is a computed variable based on student composition in primary school classes.

This paper describes the methods for producing the MED-registries and how they can be used in analysis. We present methods for using administrative school data in official statistics and for assessing the quality of such data. Finally, we present results of two published articles based on these data.

Keywords: Education, Analysis, Policy-development, Matching, Peer effects, Primary school, Teacher effects

1. Introduction

In 2019 Statistics Denmark launched the project on Matched Educational Data (MED). The project is funded by the DRDS group (Danish Research Data for the Social Sciences) and the overall vision is to develop a register that links students from primary school to upper secondary education together at class level, and connect information on the teacher and activities across the whole school year. This is something that has not been done before - in Denmark or worldwide.

In this paper we will give an overall introduction to the MED-register and the analytical possibilities that it contains. The innovative part of the project especially lies in the data on activities (school subjects) which is based upon relatively new administrative registrations at the education institutions. This gives the opportunity to link students and teachers to the individual activity. The first steps of the MED-project revolved around statistical products regarding aggregated data on activities for an entire school year which was derived from ministerial data sources. Even though much can be learned by analysing activities on an aggregated level it is now possible to link students and teachers to individual activities due to the data derived from the administrative systems at the educational institutions.

With the new register, large amounts of data will be generated as more than 80 million activities are generated exclusively from the administrative systems in primary school in the school year 2020/2021. The register is still under development and will be made available to the researchers, ministries and others in 2023 and data will be published on DST's website. In the long term, it is a success criterion that the register can become central to new policy development in the field of education and provide knowledge that can be used in the school administration across sectors.

2. Vision of the MED-register

The MED project is funded by the DRDS group which was formed as a consortium in connection with the Ministry of Education and Research's offer of research-funding. The consortium consists of six national universities, Danmarks Nationalbank, The Danish Economic Councils and the Rockwool Foundation's research unit.

It was highly welcomed by the DRDS group that resources where spend on generating matched educational data, which match students to classrooms and teachers as well as performance measures, as it would generate a dataset which the DRDS group had no knowledge of existing anywhere else. Thus, the vision is to create a crucial new register, which provides opportunities for new research at the international level.

As student output varies across schools at all education levels it is the expectation that the MED-register will provide a unique opportunity to analyse new and important research questions as the register allows researchers to deeply analyse questions about the roles of teachers and the effect of the socioeconomic composition of classes etc. These factors constitute some of the key external factors that can explain differences of outcome between different schools across the education sector. For example, Chetty, Friedman and Rockoff finds that teachers “quality” matter for student test scores, after controlling for student characteristics (Chetty, Friedman and Rockoff 2014a, 2014b, and 2016). And at a more descriptive level, The Danish Ministry of Education publishes a ranking of schools by student grades, corrected for observed characteristics, which shows that between school variation can explain almost 70 % of students’ grades, whereas the students’ socioeconomics background characteristics explains only 25 % of the variation (UVM 2017).

Many questions are still to be answered; “Are teachers’ impacts additive over time?” “What is the effect of substitute teachers in the education environments?” Questions like these, could be answered with the MED-register. The MED-register therefore provides an opportunity for producing ground-breaking work on understanding what the important input factors are in the production of knowledge among students and other dimensions of output.

3. Products based on historical (aggregated) data

As of May 2022, two matched educational data products are available at Statistics Denmark for use by analysts and researchers.

3.1 Product 1: Primary school teacher register

The first of these products is the primary school teacher registry, which provides information on the link between students and teachers for every school year[1]. While these data are not nearly as detailed as the data linking students and teachers to every activity, much can be understood by analysing activities at the aggregate level. Furthermore, the cost of producing data at the aggregate level is significantly lower, and the quality assurance of this type of data is significantly less complex.

3.1.1 Production

Production of the primary school teacher registry is achieved by taking advantage of two annual data collections already in operation for other purposes.

The first of these is the Danish Ministry of Children and Education's data of public-school teachers. The ministry collects these data for monitoring planned teaching hours as well as the share of teaching hours taking place with teachers with teaching competency in the respective school subjects.

Statistics Denmark administers the second data collection, which is the basis for the registry of primary school students in Denmark. This registry provides information on all students in primary schools in Denmark, including which school, grade and class the students are attending.

With these two data sources in hand, information on both teachers and students in Danish primary schools are available. For visualization purposes, example data of the two data sources are provided in Table 1 and Table 2 below:

[1] As of May 2022, data for every school year 2014 – 2020 are available for use.

Table 1: Statistics Denmark's primary school registry

School year	Student ID	School ID	Grade	Class
2019	Student 1	School 1	4	4. B
2019	Student 2	School 1	4	4. B
2019	Student 3	School 1	4	4. B
2019	Student 4	School 2	7	7. Y
2019	Student 5	School 2	7	7. Y

Table 2: Danish Ministry of Children and Education's data of public-school teachers

School year	Teacher ID	School ID	Grade	Class	Team	Subject	Hours	Competency*
2019	Teacher 1	School 1	4	4. B		Danish	210	Formal
2019	Teacher 2	School 1	4	4. B		Mathematics	150	Formal
2019	Teacher 3	School 1	4	4. B		English	60	Equivalent
2019	Teacher 4	School 2	7	7. Y		Geography	60	Non-formal
2019	Teacher 5	School 2	7		7. X & Y	Biology	60	Formal

* Information on teaching competency is provided by the schools' management: 1) 'formal' meaning that the teacher has studied to become a teacher in the respective subject, 2) 'equivalent' meaning that the teacher's competency is not formal but equivalent to formal and 3) 'non-formal' meaning that the teacher has neither formal nor equivalent teaching competency in the respective subject.

While there exists no explicit link[2] between students and teachers in the two data sources, it is possible to match populations via the school, grade and class. In technical terms, the matching is achieved by using a composite key consisting of the variables: 'School year', 'School ID', 'Grade' and 'Class'. By virtue of this composite key, the information provided in the product of the match pertains to singular classes per school year, which is in line with how the Danish primary school system is structured primarily.

However, the different purposes by which the two data sources are collected and administered introduces some difficulties in matching the data correctly. Chiefly among these difficulties is the fact that 'Class' is a free form text variable, which is a datatype that itself can introduce a variety of errors when used in a composite key. Furthermore, the free form text variable 'Team' exists in the Danish Ministry of Children and Education's data of public-school teachers, thus providing two different ways of registering which classes a teacher taught in a given school year.

To address the difficulties highlighted above, Statistics Denmark employ a total of 14 different text matching procedures (see appendix 1) for the 'Class' / 'Team' variable, while matching directly on the remaining variables in the composite key ('School year',

[2] A unique key by which to match the records in both data sources.

'School ID' and 'Grade'). Based on rigorous testing of the 14 individual procedures, the procedures are executed in hierarchical order by estimates of error rates (see appendix 1).

With use of this matching method, the primary school teacher registry is produced. See below for example data:

Table 3: Primary school teacher registry

School year	Student ID	Teacher ID	School ID	Grade	Class	Subject	Hours	Competency*
2019	Student 1	Teacher 1	School 1	2	2. C	Danish	165	Formal
2019	Student 1	Teacher 2	School 1	2	2. C	Danish	165	Formal
2019	Student 1	Teacher 3	School 1	2	2. C	English	30	Non-formal
2019	Student 1	Teacher 4	School 1	2	2. C	Religion	30	Equivalent
2019	Student 1	Teacher 5	School 1	2	2. C	Mathematics	150	Formal
2019	Student 1	Teacher 6	School 1	2	2. C	Nature/technology	60	Formal
2019	Student 1	Teacher 7	School 1	2	2. C	Physical education	60	Non-formal
2019	Student 1	Teacher 8	School 1	2	2. C	Music	60	Non-formal
2019	Student 1	Teacher 9	School 1	2	2. C	Arts	60	Equivalent

The primary school teacher registry is the Cartesian product[3] of Statistics Denmark's primary school registry and the Danish Ministry of Children and Education's data of public-school teachers.

This registry contains one row for every teacher a public-school student has had in every subject per school year. Furthermore, the variables 'Hours' and 'Competency' provides information on the planned teaching hours and teaching competency for each teacher.

3.1.2 Quality assurance

As stated earlier, these data are at the aggregate level. Therefore, the quality assurance of these data is far less complex compared to matched educational data at the level of activities. This is due to the fact, that the regulation of the Danish public schools is stringent in terms of planned teaching hours and compulsory subjects, which

[3] The table produces approximately 10 records per student every school year. It would be possible to store the data in two separate tables in a relational database to minimize the memory usage by reducing redundancies. However, while the amount of records per school year remains manageable, Statistics Denmark stores the Cartesian product to allow for a simpler use of data for analysis and research purposes.

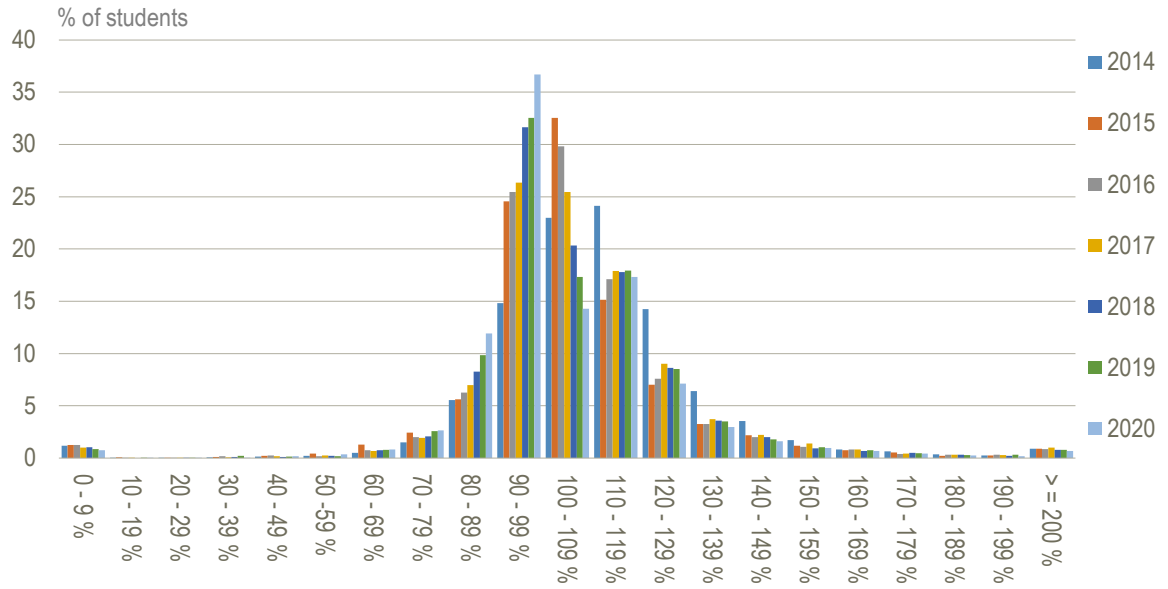
is exactly what is measured in the primary school teaching registry. As such, it is possible to use the Danish Ministry of Children and Education's regulation in this area as a guideline for quality assurance (see appendix 2).

Using this regulation, we are able to identify which subjects the students are expected to have every school year (dependent on grade), and how many teaching hours can be expected for every subject. This provides a key component on which the quality assurance of the public-school teacher registry rests.

As of May 2022, data for the school years 2014-2020 are available, and for each of these school years, the students' data form a normal distribution around 100 % in the observed versus expected sum of teaching hours as shown in Figure 1.

The latest available school year, 2020, shows that 80% of students fall within an interval of 80-119% of expected teaching hours in the registry, while 90% fall within the interval 70-129%. This is a noticeable improvement compared to the first year of data collection, 2014, where 67% of students had 80-119% of expected teaching hours in the registry and 83% had 70-129%.

Figure 1: Observed versus expected sum of teaching hours* per student 2014 - 2020



* The expected sum of teaching hours is based on the regulation of teaching hours in Danish public schools, which is administered by the Danish Ministry of Children and Education. The expected and actual sum of teaching hours is measured only in the compulsory school subjects in each grade, as stated in the regulation.

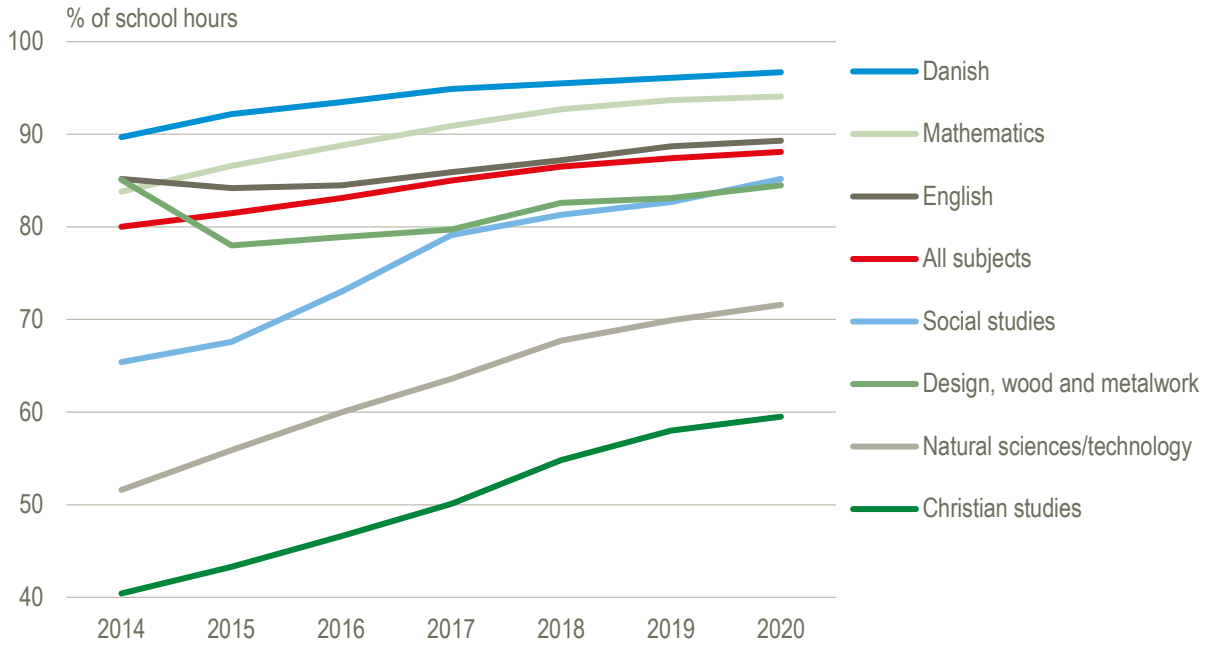
The reference period for a school year is October 1st the year prior until September 30th in the respective year.

As such, the data in the primary school teacher registry is, in 2020, closely aligned with the regulation of teaching hours in Danish public schools.

3.1.3 Preliminary results from a 2021 article

To demonstrate some of the analytical possibilities of the primary school teacher registry, Statistics Denmark published an article in 2021 (A.S. Nielsen 2021) with focus on teaching competency and teacher changes. Updated data from the 2021 article is presented below in Figure 2.

Figure 2: Planned teaching hours with formal teaching competency or equivalent

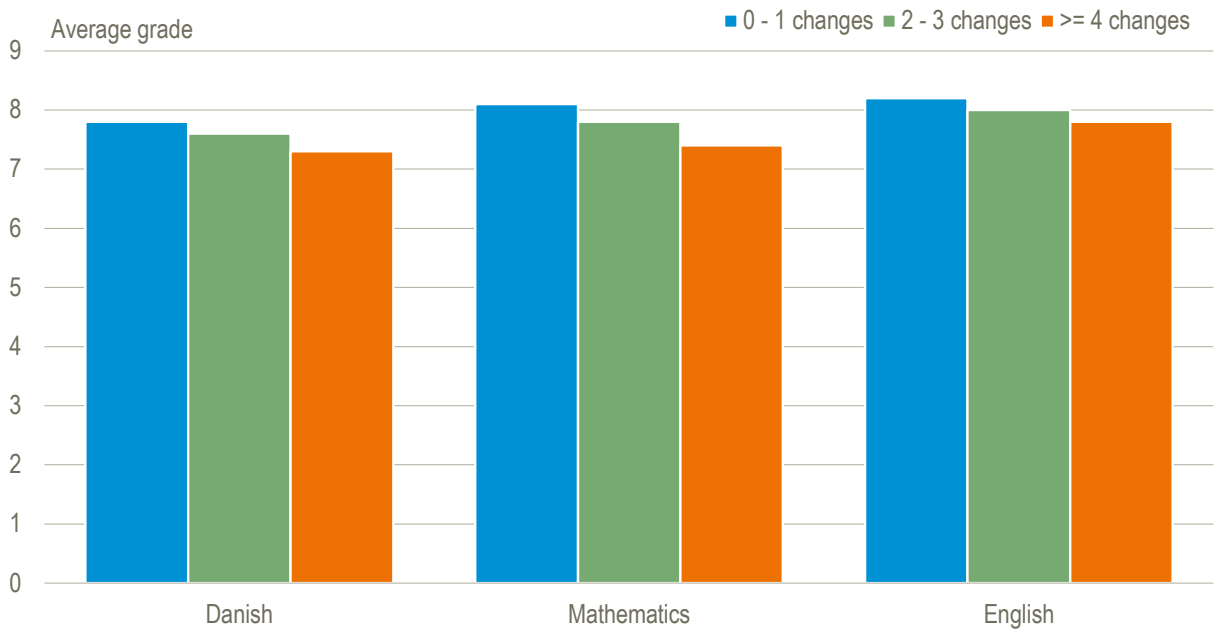


Subjects in the graph is a sample of all school subjects in Danish primary schools. For the remaining subjects, the share of school hours with formal teaching competency or equivalent rose with 2.5 to 17.9 percent points.
The reference period for a school year is October 1st the year prior until September 30th in the respective year.

From 2014 to 2020, the share of planned teaching hours with at least one teacher with formal teaching competency or equivalent rose steadily every year from 80% to 88% across all school subjects. Especially Natural sciences/Technology, Social studies and Christian studies saw a rise with respectively 20, 20 and 19 percent points in the period.

In the core subjects, Danish and Mathematics, which are compulsory at every grade, the share of planned teaching hours, formal or equivalent competency, rose to 97 and 94 % by 2020.

Figure 3: Final grades by number of teacher changes



The population for these statistics are public school students in 3rd grade by October 1, 2014. Teacher changes are calculated by comparing the students' teachers in the respective subjects across school years.

Another analysis made possible by the primary school teacher registry is the relationship between teachers and students. By looking across school years and subjects for every public-school student, it is evident that a relationship exists between the number of teacher changes a student has experienced, and their final grades in the respective subjects.

Figure 3 shows that, In the case of Danish, Mathematics and English, a negative linear relationship can be observed between the number of teacher changes and final grades – i.e. the more teacher changes a student has experienced, the lower their final grades are on average. Among these three subjects, the strongest effect can be observed for Mathematics, in which student with four or more teacher changes received a grade average 0.7 points lower than students with 0-1 teacher changes.

3.2 Product 2: Class ID

The second statistical product available as of May 2022, is the 'Class ID', which identifies groups of primary school students across time (school years) and space (schools).

Asking any Danish primary school student, which class they are in at a given moment, they would likely provide the name of their class, for example "4. B" and perhaps the

name of their school. Primary school data on classes are typically stored in the same way, i.e. with a school ID, a class name and a time period. Thus, defining class population at a given time is relatively simple, but complexity rises when the aim is to define classes across time and space.

For example, a class of primary school students might be named “4. B” one school year, and “5. Y” the next, since schools are free to name classes as they wish. Another common occurrence in the Danish primary school system is that some schools only have classes up to 6th grade, resulting in the students changing schools when they begin the 7th grade. Furthermore, the classes’ student population one year might have changed the next because of students changing classes and / or schools, which is common as well.

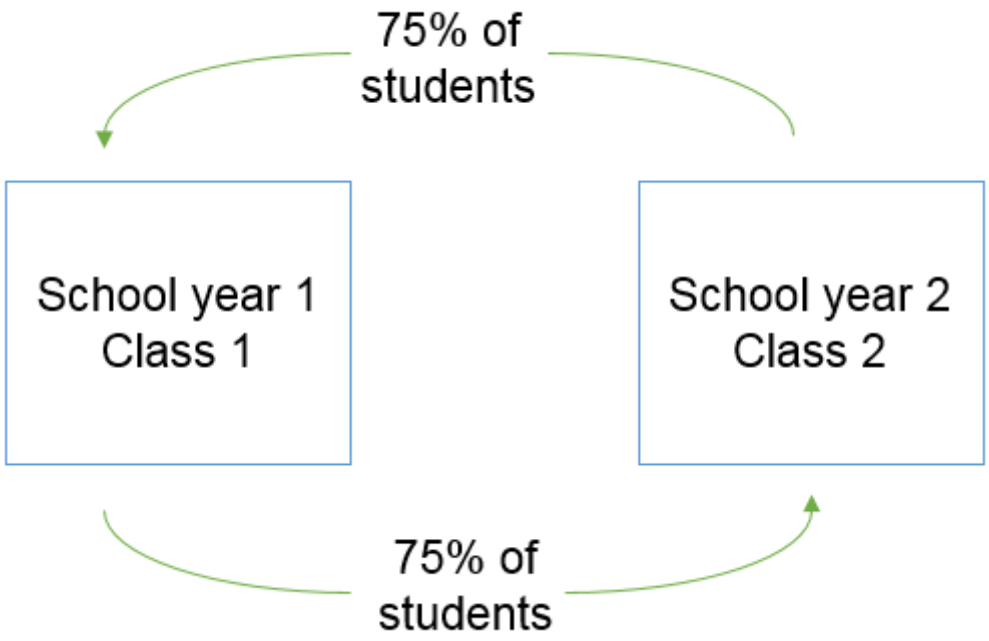
The complexity of the classes in the Danish school system provides complications when one wishes to study peer effects and other topics pertaining to primary school students at the group level longitudinally. To address these complexities, Statistics Denmark has developed the Class ID.

3.2.1 Production

Statistics Denmark’s primary school registry forms the basis for the class ID (see Table 1) and classes are defined by variables in this registry.

A class can be defined as a group of primary school students participating in the same school activities over a school year. In technical terms, the class is defined as a group of primary students who, in the same school year, are registered with the same school ID, grade and class name in the primary school registry.

A unique class ID is initially assigned to every class in a school year following the technical definition above. Then, looking at the following school year, if 75% of the students in a given class can be found in a class together the year prior, and 75% of the students in the class of the prior year can be found in the current class, the unique class ID remains unchanged. However, if these criteria are not met, a new class ID is assigned to the class. See below for an illustration of this procedure:

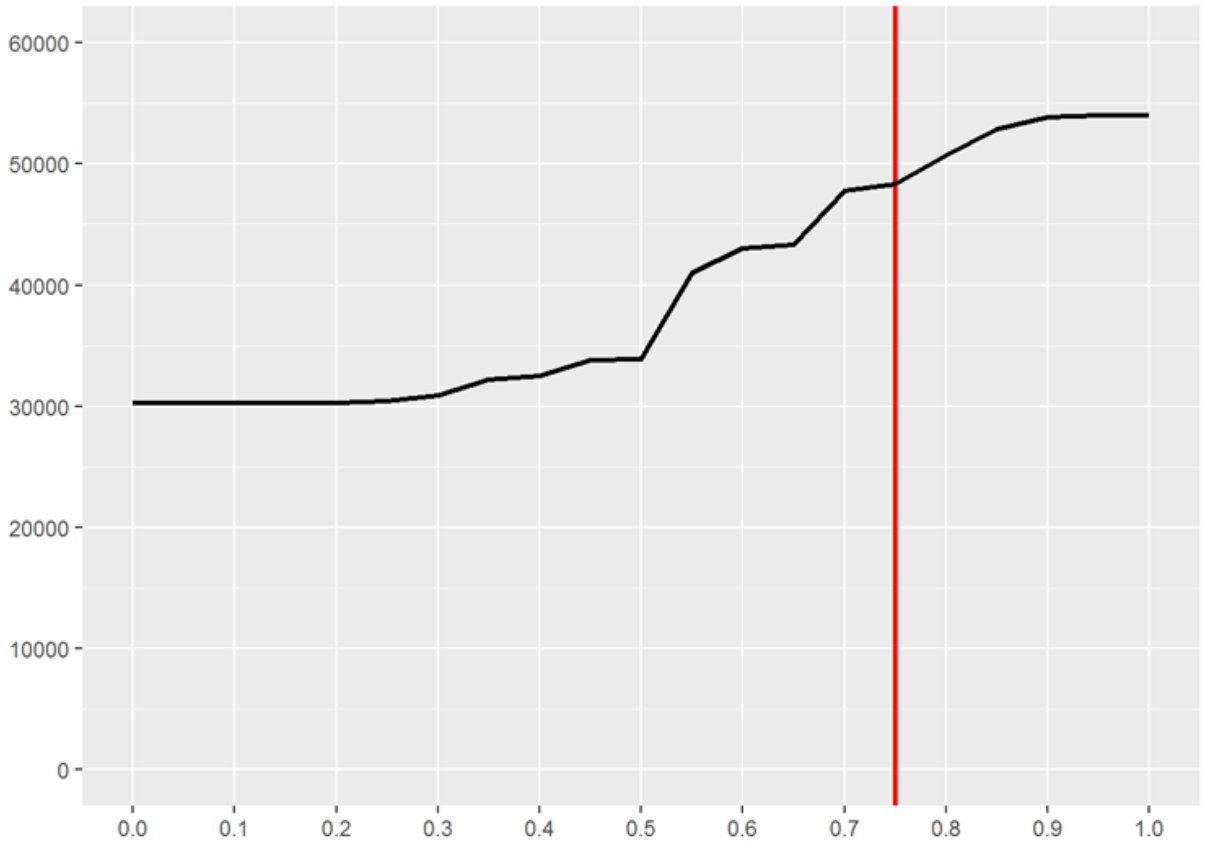


In this view, if less than 75% of the students remain in a class together from one school year to the next, the class population will have changed to such an extent, that the identity of the class has changed as well.

Prior to settling on the 75% minimum, a review of the scientific literature on school class dynamics was completed. This review revealed a scarcity of longitudinal studies of school classes and a general arbitrariness in terms of cut-off criteria for when a school class can be assumed to have the same identity across school years.

Because of the lack of research, an empirical approach was employed to find suitable criteria.

Figure 4: Number of class ID's by cut-off criteria

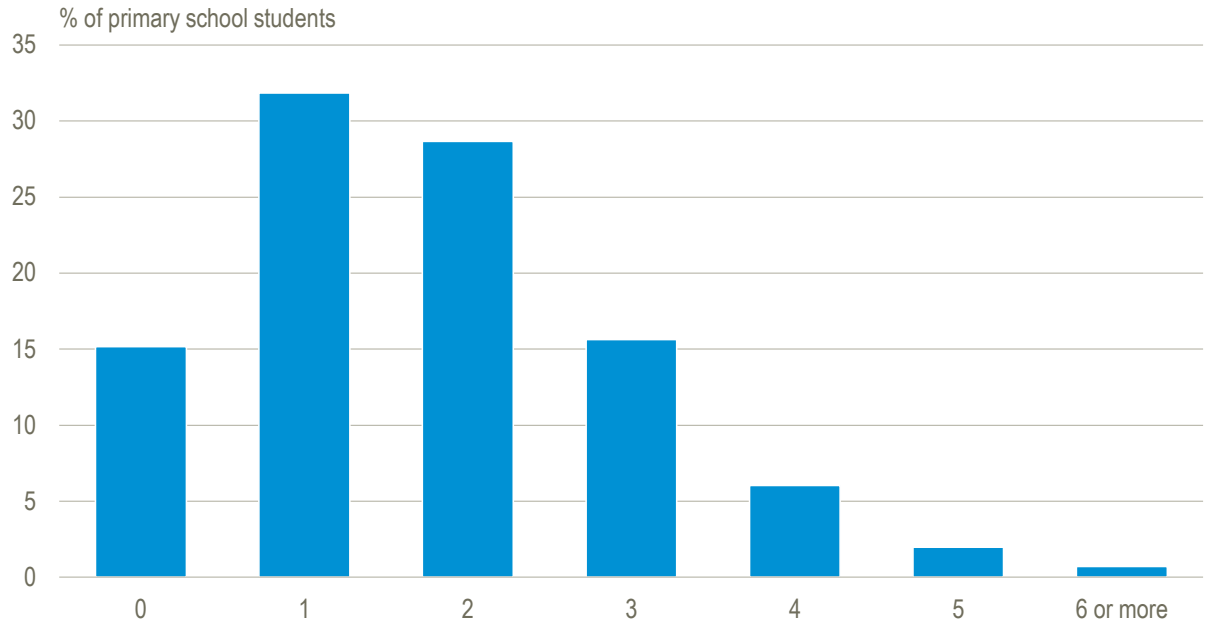


Analysis show that the amount of class ID's rise significantly at the 50% mark, and then again at the 65% and 75% mark. On this basis, the cut-off criteria was set to the highest of these break points, 75%, in Statistics Denmark's class ID. However, arbitrariness is not completely removed, and following new research in the field, it will be possible to revise the cut-off criteria in the future.

3.2.2 Results from different uses of the class ID in a 2020 article

Statistics Denmark publishes a paper in 2020 demonstrating different ways to use the class ID in analysis pertaining to primary school students (A.S. Nielsen et al. 2020). Firstly, the paper showed that, from grade 0-9 a primary school student typically changes class one or two times (see Figure 5). Furthermore, it is not uncommon to have zero or three class changes, while it is very uncommon to have more than four class changes during primary school education.

Figure 5: Primary school students by number of class changes from grade 0-9



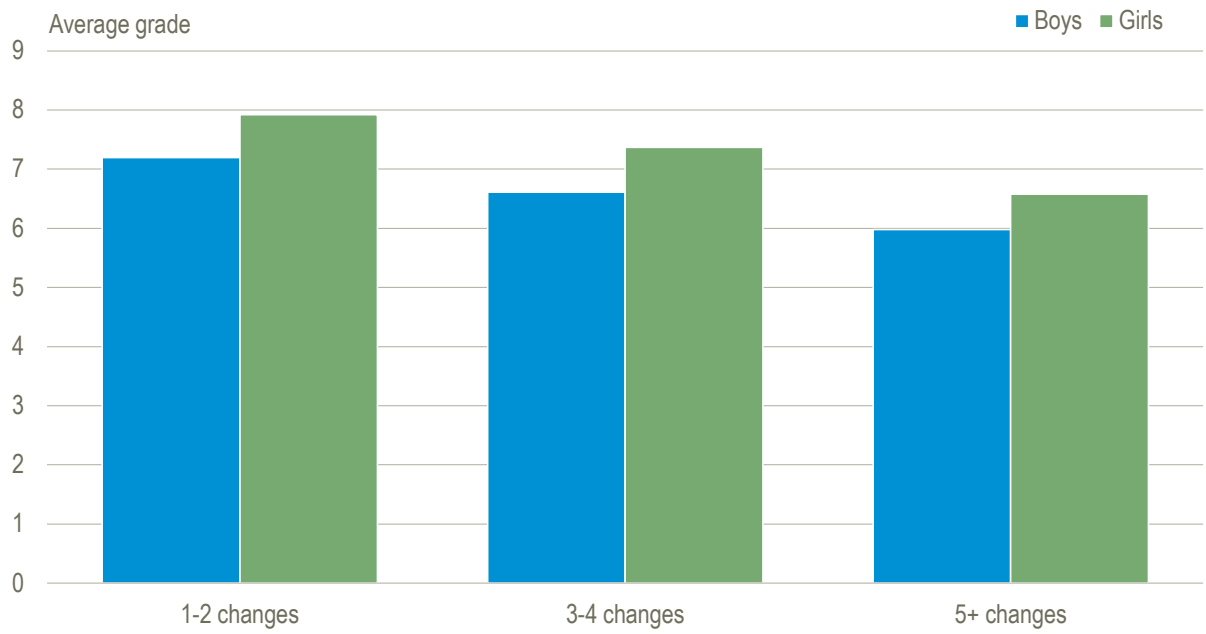
Period of reference: school years 2007 – 2019
Population: All Danish public school students who started in school (grade 0) in school years 2007, 2008 or 2009.
The reference period for a school year is October 1st the year prior until September 30th in the respective year.

Looking at possible explanations for the differences in number of class changes, we find that only 15 % of all class changes happens in school years, where students change address or family status[4]. This suggests that class changes primarily are caused by other factors.

Class changes can in itself be an explaining factor when looking at output measures such as grades. Here we find that primary school students with more class changes on average receive lower final grades on average when they finish primary school by grade 9, as illustrated in Figure 6.

[4] Statistics Denmark employs the variable 'family type' to determine family status. In this case it primarily refers to a change from parents living together to separated parents.

Figure 6: Final grades by number of class changes from grade 0-9



In conclusion, we find a broad spectrum of use cases for the class ID variable[5]. This includes in descriptive statistics as well as a dependent, independent or controlling variable in regression models.

4. Linking students and teachers to individual activities

The previous sections of the paper describes two MED-products based on aggregated and historical data. However, the main vision of the MED-register, is to create a link between teachers and students via the individual activity, which allows for more detailed analysis.

The link between teachers and students on activity level is the most comprehensive part of the project. This is because the register is based on a high-frequency data collection process which is structured in a relational database. The cost of producing data at activity level is therefore also significantly higher than producing data on aggregated level.

[5] More use cases and results are presented in A.S. Nielsen et al. 2020.

The high-frequency database is still under development at Statistics Denmark and the MED-register at activity level is therefore not yet fully developed. However in this part of the paper we will take a closer look at the data sources which are used for generating the MED-register at activity level for the primary schools and look into some of the analytical aspects that the detailed data can provide.

4.1 Production

The main data source for primary school level is derived from the *Aula-portal.* The Aula-portal is a result of an agreement from 2014 between KL (The national association of municipalities) and the Government on a digital boost for primary and lower secondary schools. The agreement meant that all primary and lower secondary schools should have a digital collaboration and communication tool that supports the work at the schools. All 98 municipalities decided to join.

The Aula-portal is used for communication between schools, students and parents and the essential part is that the portal provides an overview of the weekly schedule for the group which the individual student is affiliated. The Aula-portal therefore contains central information for producing a MED-register at activity level. The central data in the Aula-portal is the individual activities, the group connected to each activity and the ID for the teacher connected to the activity.

Figure 7: Illustration of a daily schedule in the Aula-portal

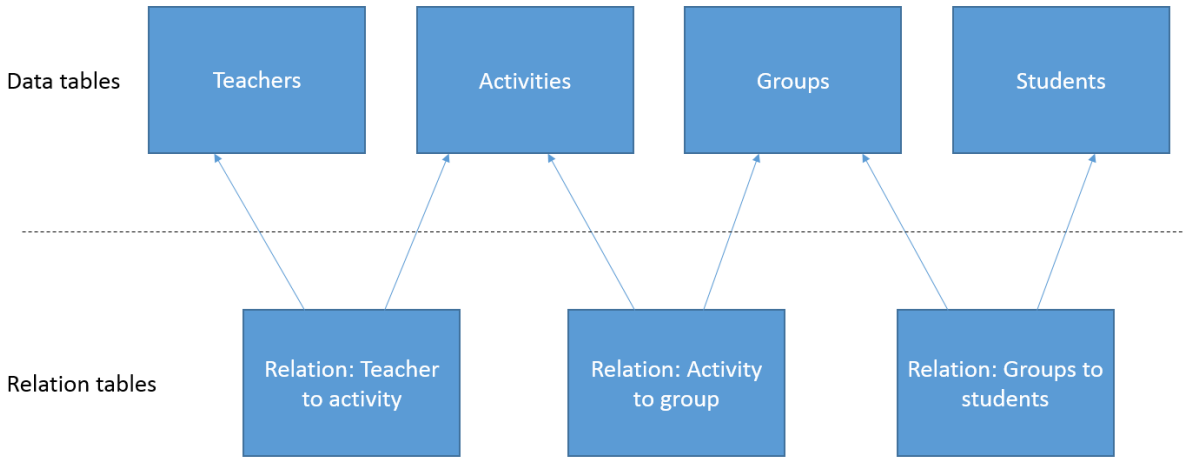
Period of time	Activity	Teacher_ID	Group
8.15 – 9.00	Danish	Teacher A	8.B
9.00 – 9.45	Danish	Teacher A	8.B
10.00 – 10.45	Mathematics	Teacher B	8.B
10.45 – 11.30	History	Teacher C	8.B
12.00 – 12.45	Sports	Teacher D	8.B + 8.A
12.45 – 13.30	Sports	Teacher D	8.B + 8.A
13.45 – 14.30	English	Teacher E	8.B

The institutions update the Aula-portal several times a day as schedules for students and teachers are frequently revised. Statistics Denmark has therefore set up a high frequency data collection process which ensure that all information on activities, groups and teacher is received. This means that Statistics Denmark receives several thousand reports every week.

The Aula-portal does not contain person specific information. Specific information on students and teachers therefore is derived from the UNI-login system which is developed by the Ministry of Education. The UNI-login system is an identification tool which is used for a large number of digital services in the field of education. Via the UNI-login system, the individual primary school can get access to a range of digital teaching aids. And the UNI-login system is also used to link students and teacher to the groups which are used in connection with the schedule display in AULA. Statistics Denmark has therefore also set up a high frequency data collection from the UNI-login system to ensure that all specific information on students and teachers is received.

Data from the Aula-portal and the UNI-login system are stored in a relational database which consists of four data tables and three relation tables which ensures the connection in the database. All in all the MED-register on activity level is therefore composed by seven tables as highlighted below.

Figure 8: Tables in the relational database

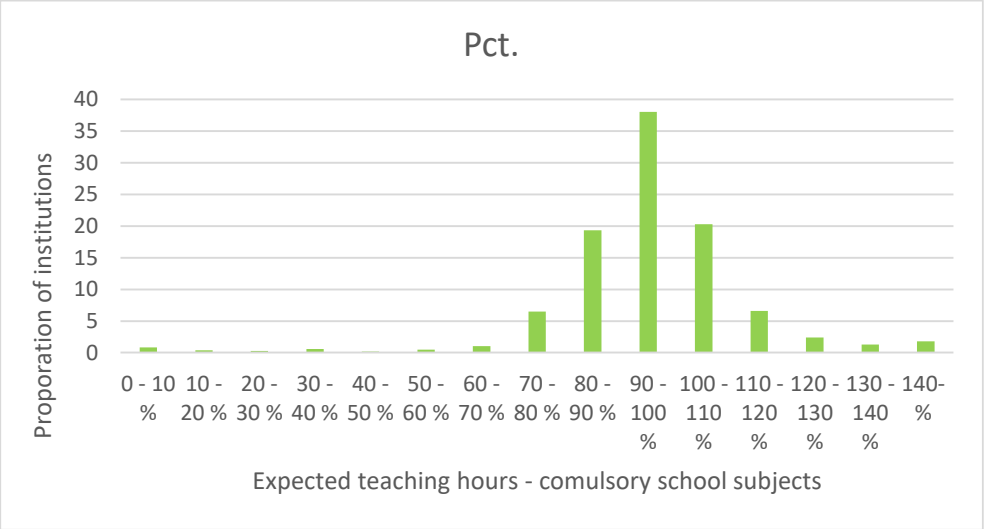


4.2 Quality assurance

The major challenge regarding the quality is that it is not possible to get the institutions to verify the data on activities. This is primarily due to the fact that the Aula-portal is a tool which is used for the daily planning and it can therefore not be expected that errors will be corrected backwards in time. In addition the amount of data which is generated via the Aula-portal is extremely large. For example, more than 80 million activities were collected in the school year 2020/2021, and it would be unreasonable to ask the institutions to verify such detailed data.

Therefore, the primary source for verifying the quality is the Danish Ministry of Children and Education's regulation on teaching hours (see appendix 2). By using the regulation, it is possible to get an insight into the quality by identifying whether the students receive the amount teaching hours that are expected.

Figure 9: Actual vs. expected sum of teaching hours* per student 2020/2021



* The expected sum of teaching hours is based on the regulation of teaching hours in Danish public schools, which is administered by the Danish Ministry of Children and Education. The expected and actual sum of teaching hours is measured only in the compulsory school subjects in each grade, as stated in the regulation.

The data for the school year 2020/2021 shows that 84% of students fall within an interval of 80-119% of expected teaching hours, while 93% fall within the interval 70-129%. Overall the data derived from Aula is, in 2020/2021, closely aligned with the regulation on teaching hours.

However, while the quality is acceptable at institution level, there are problems when the focus is directed at the micro-level. This is due to that it cannot be expected that

the weekly schedules will not have ‘gaps’ as there will be incorrect or missing registrations in the Aula-portal.

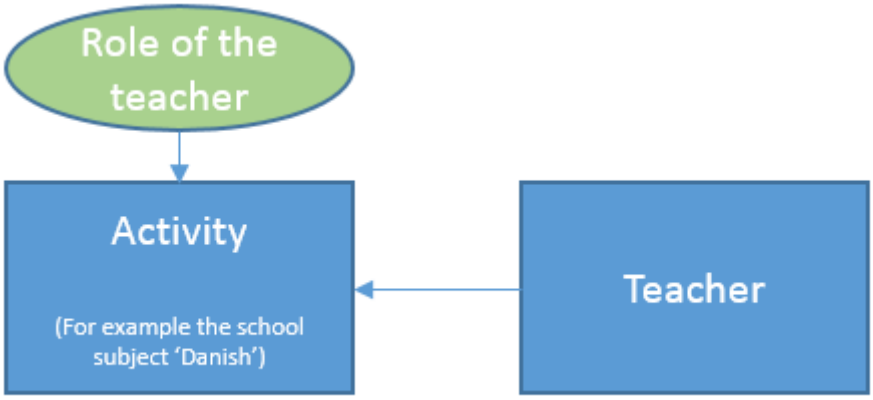
Another major challenge is that the subject of the activities is not indicated with a fixed label in the Aula-portal. On the contrary the subject is typed in manually by the teachers who maintain the schedule. This means that a subject can be specified in a way that the individual teacher prefers, and the consequence is that subjects are registered in many different ways in the Aula-portal. For example we have identified more than five thousand different designations which can be interpreted as the school subject ‘Danish’. Therefore, it has been necessary to develop a comprehensive text reproduction procedure, to ensuring that the individual activities are indicated with the correct school subject.

Finally, the amount of data poses a problem for both for quality assurance but also for the general use of the register. As previously mentioned, more than 80 million activities were collected in the school year 2021/2020. And when data on activities is combined with the data on groups, teachers and students, the efficiency of the relational database drops drastically. This meant that it took four to five working days to generate a quality report at the beginning of the development process. Therefore, many resources have been invested in optimizing the database structure, to ensure that it works efficiently.

4.3 Analytical possibilities

Even though the MED-register on activity level is still under development it is possible to identify some of the analytical possibilities that the register contains. For example, the register provides an opportunity to examine the use of substitute teachers as the role of the teacher linked is to the individual activity as illustrated below.

Figure 10: Role of the teacher in the database



The information on substitute teachers can be broken down to specific institutions, classes or school subjects and can also be combined with data on grades. The register therefore provides a opportunity to examine the amount of activities which is carried out by substitute teachers and can help shed light on whether substitute teachers is more prevalent in certain school subjects and institutions etc.

As the register contain all individual activities it also provides the opportunity to dive into the composition of the school day for any given group. One could imagine that such an analysis could help identify how the school day could be put together in the most appropriate way, to ensure that students get the most out of the teaching.

Figure 11: Example of a weekly schedule for a specific group at primary school

Period of time	Monday	Tuesday	Wednesday	Thursday	Friday
07.45 – 08.30	Danish	Craft and design	Mathematics	Mathematics	Danish
08.30 – 09.15	Religion	Craft and design	General support	English	Danish
09.45 – 10.30	Mathematics	Music	Danish	Danish	Danish
10.30 – 11.15	Danish	Mathematics	Music	Danish	Technology
11.45 – 12.30	Visual arts	Mathematics	Danish	Sports	Technology
12.30 – 13.15	Visual arts	History	English	Sports	General support

Due to the level of detail, the MED register provides an opportunity to answer many different research questions. The new register will be published at the end of 2023,

and it is the expectation that the register can become central to new policy development in the field of education.

5. References

Nielsen, A.S, Nielsen, K. M. R. & Andersen, A. K. (2020). Folkeskoleelever med mange klasseskift får lavere karakterer ved afgangsprøven. www.dst.dk/da/Statistik/nyheder-analyser-publ/Analyser/visanalyse?cid=40003

Nielsen, A.S (2021). Flere timer med uddannede lærere i folkeskolen. www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=46850

Statistics Denmark

UVM (2017). Socioøkonomisk reference for grundskolekarakterer. https://uvm.dk/statistik/grundskolen/karakterer-og-test/sociooekonomisk-reference-for-grundskolekarakterer

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a). "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." American Eco-nomic Review 104.9: 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b). "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." Amer-ican Economic Review 104.9: 2633-79.

Chetty, Raj, John N. Friedman, and Jonah Rockoff (2016). "Using lagged outcomes to evaluate bias in value-added models." American Economic Review 106.5: 393-99.

6. Appendices

Appendix 1: Matching rules

To test the reliability of the 14 matching rules, we take an explorative approach and employ a loop over the 14 rules and validate each row by four criteria:

1. If rule matches 1 subject and one teacher, the match is considered valid

2. If rule matches the subject 'Physical education' and four teachers or less per class, the match is considered valid
3. If criteria 1 and 2 are not met, but the sum of teaching hours for the subject is within a margin of 10 % compared to the minimum hours stated in the Danish Ministry of Children and Education's regulation, the match is considered valid
4. If criteria 1, 2 and 3 are not met, but a maximum of 3 teachers are matched per class and the sum of teaching hours is a maximum of (number of teachers * minimum teaching hours * 1.1)

If a row is matched and not validated by at least one of the four criteria, the row is send to manual inspection.

Matching rules for Statistics Denmark's primary school teacher registry

Number	Procedure	Example	Error rate*
1	1-1 match by class name	'7B' matched as is	0,02
2	Removal of symbols and whitespaces	'1-A' converted to '1A'	0,03
3	Removal of leading zeroes	'09B' converted to '9B'	0,02
4	Removal of redundant letters	'M6B' converted to '6B'	0,04
5	Removal of redundant letters and whitespaces	'M 4 C' converted to '4C'	0,01
6	Partitioning of aggregated class names by grade	'1234A' partitioned to '1A', '2A', '3A' and '4A'	0,19
7	Partitioning of aggregated class names by name	'6ABC' partitioned to '6A', 6B' and '6C'	0,2
8	Partitioning of aggregated class names by delimiter	'8A/8B/8C' partitioned to '8A', '8B', '8C'	0,16
9	Removal of redundant trailing numbers, then partition	'4A4B1495091328' partitioned to '4A' and '4B'.	0,24
10	Partitioning of class names by interval	'1-4A' partitioned to '1A', '2A', '3A' and '4A'	0,20
11	Grade matched on individual class names	'Grade 7' matched on '7A' and '7B'	0,34
12	Interval of grades matched on individual class names	'Grade 8-9' matched on '8A' and '9B'	0,35
13	Concatenation of class letter and grade variables becomes matching key	'Grade 7' and 'A' concatenated to '7A'	0,47
14	If only one class exists at grade 0, grade becomes matching key	'Grade 0' matches all teachers on grade 0	0,56

* The 'error rate' is in this instance the proportion of rows send to manual inspection for each matching rule.

Appendix 2: Teaching hours regulation

The Primary and Lower Secondary School Act stipulates a minimum number of teaching hours for the school subjects for each grade[6]. The table below indicates the minimum number of teaching hours for the individual grade and school subjects as well as the minimum duration of the teaching time at the individual grades for the school year 2020/2021.

Timetal (minimumstimetal og vejledende timetal) for fagene i folkeskolen - skoleåret 2020/2021

BØRNE- OG UNDERVISNINGSMINISTERIET

Klassetrin	Bh.	1.	2.	3.	4.	5.	6.	7.	8.	9.	Timetal i alt
Humanistiske fag											
Dansk		330	330	270	210	210	210	210	210	210	2.190
Engelsk		30	30	60	60	90	90	90	90	90	630
Tysk/fransk						60	60	90	90	90	390
Historie				30	60	60	60	60	60	60	390
Kristendomskundskab		60	30	30	30	30	60	30		30	300
Samfundsfag									60	60	120
Naturfag											
Matematik		150	150	150	150	150	150	150	150	150	1.350
Natur/teknologi		30	60	60	90	60	60				360
Geografi								60	30	30	120
Biologi								60	60	30	150
Fysik/kemi								60	60	90	210
Praktiske/musiske fag											
Idræt		60	60	60	60	60	60	60	60	60	540
Musik		60	60	60	60	60	30				330
Billedkunst		30	60	60	60	30	30				270
Håndværk og design samt madkundskab (gl. lov)					90	120	120	60			390
Håndværk og design				60							60
Valgfag											
Valgfag (obligatorisk praktisk/musisk på 7. og 8. klassetrin)								60	60	60	180
Årligt timetal (minimum)*	600	750	780	840	870	930	930	990	960	960	8.610
Understøttende undervisning og pausetid	510	360	330	270	450	390	390	410	440	440	3.990
Årlig undervisningstid (minimum)	1.110	1.110	1.110	1.110	1.320	1.320	1.320	1.400	1.400	1.400	12.600

* Årligt minimumstimetal pr. klassetrin, 7. og 8. klassetrin: Det årlige minimumstimetal pr. klassetrin reduceres med 30 undervisningstimer på det klassetrin, hvor konfirmationsforberedelsen finder sted. Hvis konfirmationsforberedelsen finder sted på 8. klassetrin og det er besluttet at anvende § 16 d, stk. 2, kan yderligere op til 30 undervisningstimer af det årlige minimumstimetal pr. klassetrin flyttes til 7. klassetrin.

[6] Regulation of teaching hours in Danish public schools: https://www.uvm.dk/folkeskolen/fag-timetal-og-overgange/timetal