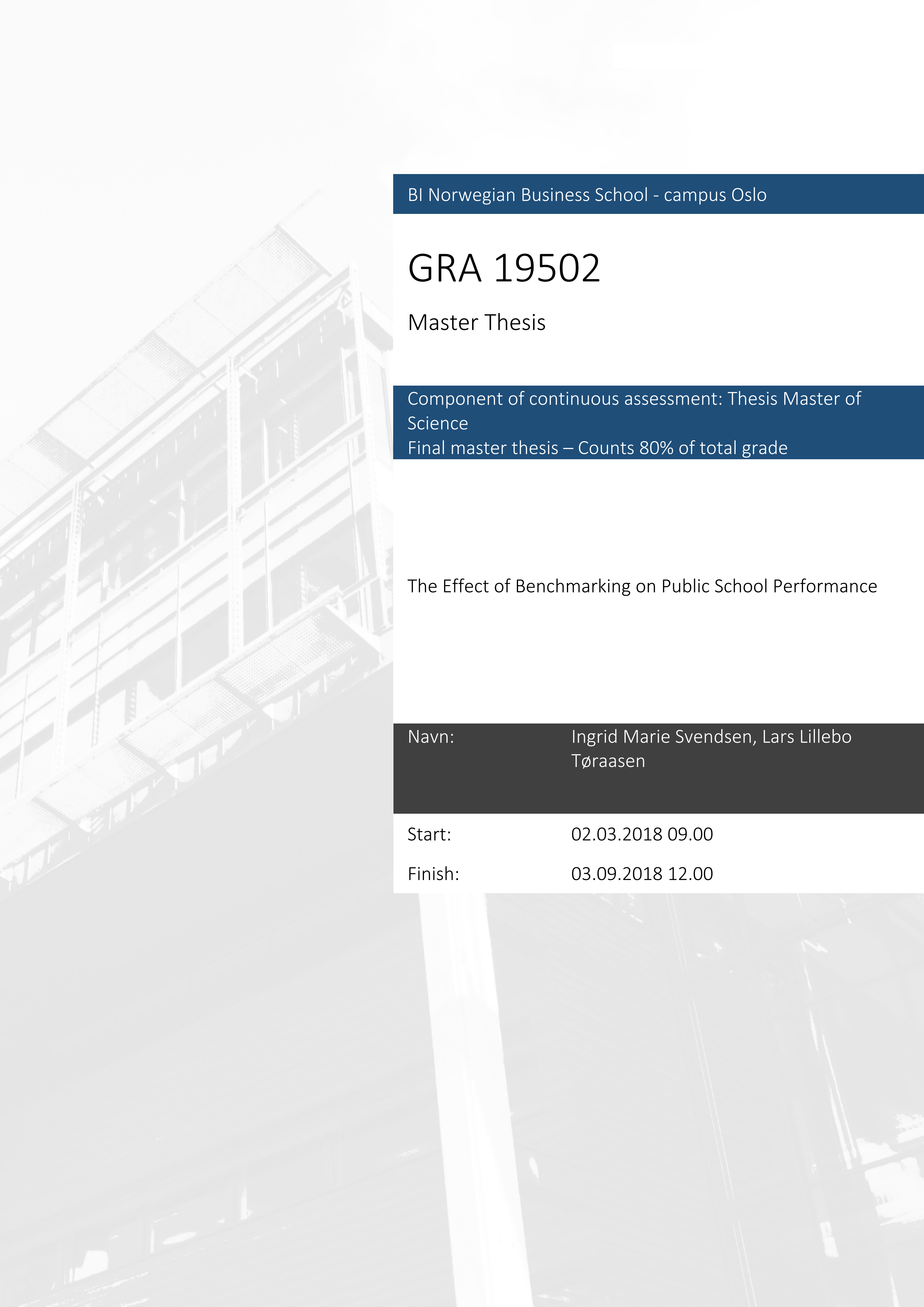


BI Norwegian Business School - campus Oslo

GRA 19502

Master Thesis

Component of continuous assessment: Thesis Master of Science
Final master thesis – Counts 80% of total grade

The Effect of Benchmarking on Public School Performance

Navn:	Ingrid Marie Svendsen, Lars Lillebo Tøraasen
Start:	02.03.2018 09.00
Finish:	03.09.2018 12.00

Lars Lillebo Tøraasen

Ingrid Marie Svendsen

The Effect of Benchmarking on Public School Performance

Date of submission:

29.08.2018

Programme:

Master of Science with Major in Economics

"This thesis is a part of the MSc programme at BI Norwegian Business School. The school takes no responsibility for the methods used, results found and conclusions drawn."

Abstract

Theory on public policy processes together with a range of private sector evidence suggest that implementing benchmarking improves efficiency, and thereby may work to improve public policy outcomes. A similar prediction is further confirmed by basic economic theory on competition exposure.

The present thesis provides an empirical analysis of this relation by studying the relationship between the implementation of benchmarking and educational performance in primary- and lower secondary education in Norwegian local governments.

The Norwegian educational sector has been subject to increased stress on performance and results. The introduction of national tests adds to this in terms of providing educational authorities, principals and local councils with information on the average achievements in 5th and 8th grade of compulsory schooling. However, our main results suggest that using these achievement-based scores to benchmark one's own school(s) does not work to improve educational performance.

Acknowledgements

This master thesis is the final part of the Master of Science in Business, major in Economics at BI Norwegian Business School. The thesis was completed during the spring semester of 2018.

We would like to extend our sincerest gratitude towards our supervisor, professor Rune J. Sørensen at BI Norwegian Business School. His supervision and constructive feedback during this process has been very helpful. In addition to this, we are also very grateful for the additional supervision provided to us from professor Benny Geys, also at BI Norwegian Business School. Their supervision has encouraged and directed the work towards completion.

Besides working on the thesis, we have both worked part-time as student assistants at the Institute of Economics at BI Norwegian Business School. This has given us experience of working with data as well as developed our programming skills in STATA.

We are also very grateful to our fellow classmates in the class of Economics for productive discussions and exchanging of ideas, while working on this thesis.

Lastly, we also thank our families and friends for support and continuous encouragement throughout our years of study.

Lars Lillebo Tøraasen & Ingrid Marie Svendsen

Content

Section 1 - Introduction

An ongoing debate in the Norwegian public administration is how to improve public sector service provisions in terms of performance. Recent years have been characterized by ongoing reforms of the municipal landscape as well as restructuring initiatives of service provisions. Local councils are exposing tasks to competition and introducing market based mechanisms at an increasing rate. In such context, benchmarking refers to a policy measure intended to enhance competition, centred around the comparisons of one's performance relative to its peers. Such policy measure raises important questions in terms of how the introduction of performance measures affect the outcome of public provisions.

Establishing a national system for quality assessment is said to represent a shift in Norwegian school governance. The idea of gathering registry data on school performance was introduced in 2004 as one of several tools forming a national education policy for quality assurance (Hovedhaugen et al. 2017). National tests are run each year in 5th and 8th grade respectively, focusing on core academic skills in numeracy, literacy and English. Educational authorities are provided with information on student achievements in order to gain insights about the general student competency at several levels of compulsory schooling. Moreover, registry data on school performance is used as a comparison at the national level to investigate and identify possible quality differences among students, schools and municipalities.

One expects the Norwegian educational system to provide an organized and structured learning methodology of the highest quality. Although benchmarking is less used within the public sector services to stimulate performance, recent practice shows an increasing trend in this regard. As an illustration, 26 percent of Norwegian local governments benchmarked educational results in 2008. By 2012, benchmarking was implemented in 63 percent of the municipalities within the educational sector (NIBR 2012).

Based on evidence from the private sector (Greve 2007; Greve 1998), there is reason to believe that the introduction of benchmarking on the municipal level influences performance scores. Therefore, this thesis is based on the hypothesis that the implementation of benchmarking as a measurement of performance in Norwegian primary- and upper secondary schools, likewise may work to cause improved educational results.

To investigate the relationship between benchmarking and school performance, the analysis consists of multiple regression models of 429 Norwegian local governments (2012 municipal structure) over the period 2004-2015. Given that we are dealing with panel data that change over time, we will be using the method of fixed effects regression. By doing so, we aim to keep the individual municipality and time effects constant. Hence, we then isolate and examine the effect of benchmarking on educational performance.

1.1 Motivation and Research Question

Recent years have shown an increasing trend when it comes to measuring and evaluating the level of performance of public sector services. Norway as a country is characterized by a strong and substantial welfare state, where the public sector is responsible for a wide range of service provisions to the population. Naturally, one would look to make sure that this sector runs as efficiently as possible, given its mandate. To prepare for future challenges, public sector re-optimization has to be looked into.

The Norwegian educational system faces major challenges in terms of academic achievements, social differences and increased dropout rates during high school. To address these challenges, the Norwegian educational system follows international management trends on quality assessment of educational policy (Roald 2010). The national test scores serve as indicators of how schools are performing based on average student achievements within school or municipality boundaries. The test score results are thereby used to monitor student performance, and to identify if adjustments are needed. However, some would argue that the national test scores only measure a limited part of a school's areas of responsibility. Equally important as satisfactory achievements are the ideas of social competence and inclusion.

The motivation of the thesis is to look at the effect of benchmarking on school performance for students enrolled in Norwegian public schools, in primary- and lower secondary school. Benchmarking will in this case serve as an incentive to improve performance by comparing educational results among municipalities and schools. Indeed, applying business-related definitions, benchmarking is defined as measuring the quality of an organization's programs, products, policies, strategies, and then comparing them to standard measurements or similar measurements of its peers. The objective in many cases is to determine what and where improvements are called for, to analyse how similar organizations achieve their performance levels, and to use information to improve own performance levels (Greve 2007).

Performance evaluations such as benchmarking, have been used extensively in private-sector businesses and corporations throughout the years as a motivational factor. The topic has been covered both in business administration theory, as well as in practice (Greve 2003, Greve 2007). Rather than looking at the impact of benchmarking on corporate performance, we analyse the impact on public administration performance in terms of educational performance in Norway. We therefore aim to analyse the following research question:

- *"Does the introduction of benchmarking in the education sector by Norwegian municipalities in 2004 have a positive effect on educational results achieved in schools within its boundaries?"*

Recent years have witnessed an increase in terms of the number of municipalities that have chosen to implement benchmarking to supervise their performance on the provision of public goods (NIBR, 2016). We aim to look at how this has affected the educational results using information from publications of *Kommunal Organisering* in 2004, 2008 and 2012. The main objective is to identify any effect of benchmarking on school performance. To a large extent, it is fair to assume that the goal of the public sector is to implement an educational policy that maximizes performance of most students. In this context, that performance level is reported through the national test scores via the Norwegian Directorate for Education and Training (Utdanningsdirektoratet). If the analysis shows that benchmarking has an effect on national test scores, it could mean that the implementation of it may improve high school dropout-rates followed by improvements in the human capital stock of the future.

Section 2 - Theoretical analysis

Public sector performance management has been a source of discussion ever since the New Public Management ideas and reforms were introduced in the 1980s, with the purpose of improved organizational effectiveness in public sector organizations (Sørensen and Geys, 2018). The fundamental question thereby – from an academic perspective as well as from the perspective of policy makers – is whether performance management systems are associated with improved performance in public organizations. Performance-based management systems may in a 'worst-case' scenario be seen as a "trend" or "fad" without real benefits in terms of organizational effectiveness. Although a stress on improving organizational effectiveness has been present for years, a meta-analysis conducted using 2188 effects in 49 studies finds that performance management has a small but positive effect on performance in public organizations (Gerrish 2016). However, the study further highlights that the impact of performance management systems increases substantially when indicators of best-practices are included, indicating that management practices have a significant impact on the effectiveness of performance management systems. When it comes to benchmarking as "a test on the influence of management practices on performance" (Gerrish 2016: 48), the study finds that benchmarking in particular appears to be an effective method to improve performance. One possible explanation is that the ability to compare own performance relative to similar organizations thereby serves as a method allowing the adoption of approaches that are known to be tied to better performance (Sørensen and Geys, 2018).

The findings above are consistent with the theory on public performance management. Such theory argues that "a central motivation behind the increasing stress on performance in public sector organizations is to help bureaucrats and elected officials make more informed decisions" (Moynihan 2008; Nielsen and Baekgaard 2015). Furthermore, Baekgaard and Serritzlew (2016) argue that performance management is introduced with the purpose to make informed decisions by presenting unambiguous information about performance of organizations. Additionally, implementing performance management systems leads to improved accountability (Sørensen and Geys 2018; Moyhihan 2008). In a range of studies, scholars indeed argue that the availability of performance data may be interpreted as a way of keeping the incumbent government accountable (James and John 2007; Boyne et al. 2009; James and Moseley 2014).

These previous findings as well as insights from performance management theory leads us to the following hypothesis on whether and how benchmarking affects educational results:

Main hypothesis

- Implementing benchmarking has a positive effect on school performance.

Current literature on benchmarking in public sector services is related to a range of public policy reports as well as recent experimental work (NIBR 2004; 2008; 20012, Gerrish 2016, Sørensen & Geys 2018). In order to compare results among groups, one may find it reasonable to identify a reference group. Greve (2007) argues that the implicit goal is to achieve as good results as the average in the reference group. Furthermore, Greve (2007) extensively discusses the use of performance measurements, and how achieving a goal may also include acceptance of risk when aiming for improvements. Additionally, besides defining a reference group, one may find it useful to compare oneself to earlier performance. This may serve as a goal to improvement as well.

Another study by Greve (1998) examines how decision makers interpret organizational performance by comparing historical and social aspiration levels. The meaning of the term aspiration is the individual's level of ambition in a given exercise. The benchmarking literature often refers to aspiration levels when aiming to achieve a common goal. Greve (1998) argues that historical performance may be used when determining the likelihood of future success during organizational changes. A common element in the benchmarking literature considers future changes, desired improvements and such that involve a significant amount of risk. That is, when aiming for improved results, one needs some kind of input that is necessary to make the desired change. In terms of our hypothesis, the method of benchmarking may be referred to as a reference when aiming to improve school performance based on own historical results. As we are comparing municipalities with themselves and to each other, the decision of implementing benchmarking may involve risk that affects the desired result. Risk may occur when aiming for better test score results through willingness to change, or because publicly available test scores affect municipal and school reputation. This is also noted as one of the key factors to motivation behind implementing measurement tools such as benchmarking in Norwegian local governments. Basic economic theory refers to increased competition as a key indicator to increased results, improved quality and so on (Grønn 2008). Hence, adopting competitive aspects among municipalities supports our main hypothesis.

We should note, however, that although municipalities implement benchmarking at a given point in time, one may not observe the hypothesised positive effect immediately. We therefore not only look at potential contemporaneous effects in our analysis below, but also consider additional research on whether benchmarking has a *lagged* effect on school performance. The central underlying reason is that public policy reforms usually take time to implement and affect policy outcomes. We therefore believe that the educational sector may respond to the implementation of benchmarking possibly only sometime after its launch.

Section 3 – Institutional Context and Data

3.1 Institutional context

We base our analysis on multiple publicly available sources covering Norwegian local governments. The institutional setting is the Norwegian political system that includes three levels; a central government, 19 county governments and 429 municipalities (given the municipal structure in 2012). The Norwegian political system is a system where the government governs until there is no longer trust by the majority in the parliament. All elected representatives are elected in periods of four years, both on national and local level, with an interval of two years between the respective elections. The political system serves as a representative democracy, where the local council is the main legislative body of the municipal government with responsibility for all aspects of the municipality's activity as well as the local budget (Borge, Falch and Tovemo 2008, 484).

Important for our purposes, local governments in Norway have a high degree of responsibility. The local governments are responsible for health care, primary schools, local roads, water and sanitation among other provisions. In terms of being local jurisdictions, they also take care of land - and regulation planning, exemptions, grants and proceedings related to private issues, as well as being in charge of local NGOs. This analysis particularly highlights local government as the provider of education at the primary and lower secondary level.

The local government level is important in many aspects. It employed about 20 percent of total workforce in 2016, and may be seen as a part of an integrated public sector where counties and municipalities are jointly responsible for implementing national welfare policies, including primary and lower secondary education (SSB 2016). Education is the second largest service sector, after elderly care. The local governments are responsible for nearly all 2848 primary and lower secondary schools through public ownership (SSB 2017a). Furthermore, another aspect is the low share of students that are enrolled in private schools. Only 3.7

percent of all students attend private schools, while about 9 percent of the schools are characterized as private, non-profit schools (SSB 2017a).

The educational field is subject to extensive and standardized regulations such as a core curriculum defined by central authorities, hours of teaching offered as well as a minimum standard of teacher qualifications. However, local authorities retain substantial autonomy in terms of developing educational policies within the structure of the national educational framework (Sørensen and Geys, 2018). This may include budgetary funds for specific educational purposes. In general, local authorities have full flexibility to manage the educational administration. As such, most local governments have an administrative position as the 'head of education' in order to maintain quality in line with the Educational Act, as well as to ensure a satisfactory learning environment and a high learning outcome. That being said, the local councils have extensive power to introduce policy reforms whenever desired (Sørensen and Geys, 2018).

Financing the educational sector

Both public and private schools are completely tax-financed (Ministry of Knowledge and Education 2011). There is, however, a discussion whether the school-finances are fairly distributed within the educational field. A study from Israel suggests that fairness and efficiency can be achieved within the same financing system, where primary school resources should be distributed based on socioeconomic conditions such as parents' education, number of siblings, immigration status as well as the socioeconomic conditions in the local community. Although there are great differences between Norway and Israel in terms of the educational system, as well the society as a whole, the study is relatable due to its innovative suggestions regarding the finances (Ministry of Knowledge and Education 2011).

The allocation criteria to Norwegian primary school do not take into account any socioeconomic conditions. The financing system rather focuses on structural differences in the local government budgets such as tax income. Researchers in the educational field argue extensively that the students' socioeconomic background explains a substantial part of the variation in students' performance, at the same time as centralized student masses with low socioeconomic background highly influence the level of costs in their respective school communities. These are conditions that may explain variation in student performance that is not covered, and thereby not controlled for, in the Norwegian educational financing system (Ministry of Knowledge and Education 2011).

3.2 Data

Our empirical analysis is conducted using elements from multiple sources of available data. We combine data on benchmarking with detailed registry data on school performance at the municipal level. We further include municipality characteristics on social and economic conditions. Lastly, we obtain municipality-level test performance indicators as an alternative measure on student performance.

Benchmarking

Major restructuring initiatives within public sector organizations have been taking place in recent years, and a wide range of studies show how local and county councils are exposing tasks to competition and introducing market based mechanisms at an increasing rate. Up until 2012, a growing number of municipalities were applying various competitive practices in different service sectors (NIBR 2012). As noted, benchmarking serves as one of these mechanisms and can be referred to as a measure of increased competition exposure on the supply side (NIBR 2004). Benchmarking is often used within personal services such as the educational sector.

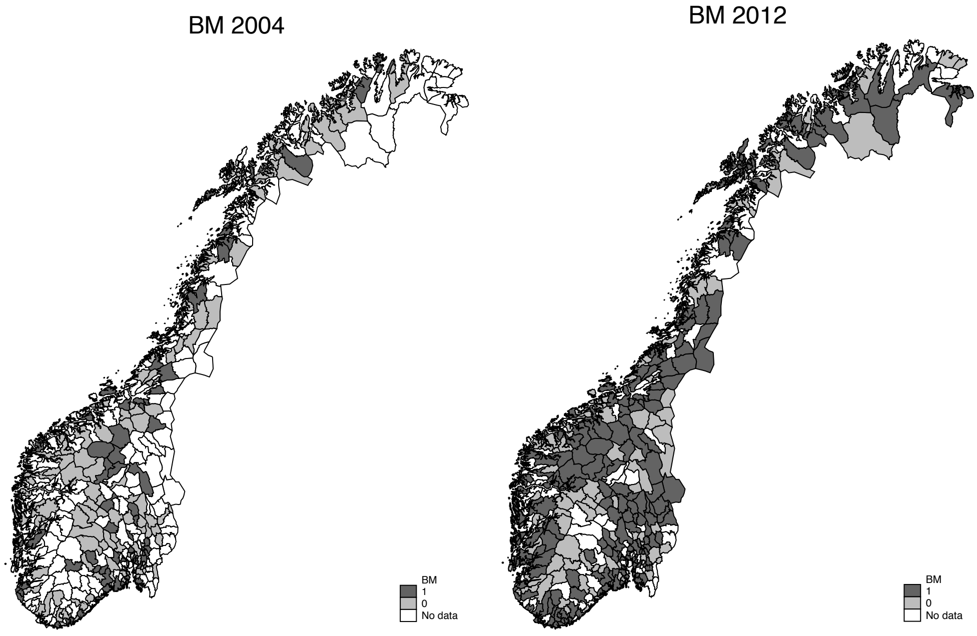
A key component in the analysis is how Norwegian municipalities to an increasing extent have introduced benchmarking in the educational sector, as well as the reasoning behind the decision of doing so. We base our analysis on three main reports; Kommunal Organisering 2004, 2008, and 2012, issued by the Ministry of Local Government and Modernization through the Norwegian Social

Science Data (NSD). Hence, benchmarking is reported every fourth year in our analysis.

The maps displayed in figure 1 indicate the development of using benchmarking within the educational sector in Norwegian municipalities in 2004 and 2012. The development may be characterized as substantial due to the increase in the use of benchmarking. Unlike the years of 2004 and 2008, 2012 represent a shift in terms of benchmarking usage within the educational sector.

Figure 1

Benchmarking implementation across municipalities



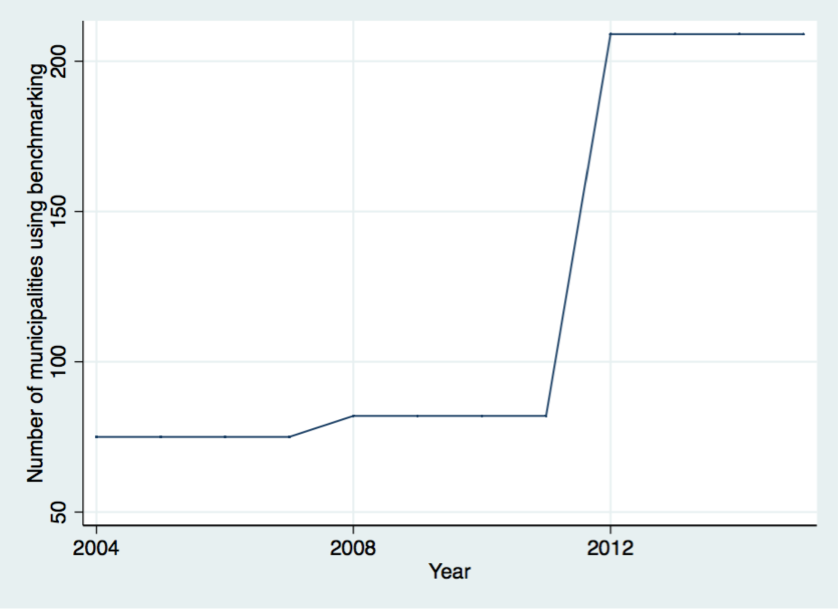
The figure shows the development in benchmarking within the educational sector across Norwegian municipalities from 2004 to 2012. Darker shaded areas imply benchmarking.

We observe that all of the larger populated areas located along the coastline are associated with an implementation of benchmarking by 2012. We also notice that there are fewer municipalities with missing data. The literature does not touch upon the reason behind this substantial change. As the national tests were first completed in 2007, educational results were most likely reported through final exam scores (10th grade) that time. There might be a connection of the increased usage of benchmarking as a result of greater availability of school performance measures. Before 2007, final exam scores in 10th grade were the only available measure on school performance that were standardized and equal among all municipalities. The introduction of the national tests in 2007 made it possible to report and compare results across schools and municipalities at an earlier stage. Therefore, the substantial increase of benchmarking is likely to be explained by the introduction of new school performance measures.

When it comes to the overall development in benchmarking, figure 2 illustrates the development in the usage of benchmarking in the educational sector in 2004, 2008 and 2012. Note that the implementation of benchmarking increased substantially over time, particularly between 2008 and 2012. More specifically, benchmarking was implemented in 23 percent of all municipalities in 2004, 26 percent in 2008, and 63 percent in 2012.

Figure 2

Overall benchmarking development



The figure illustrates the overall benchmarking implementation over the time-period 2004-2012 in Norwegian municipalities.

Registry data on school performance

We collect registry data on school performance from Skoleporten.no. These are provided through the Norwegian Directorate for Education and Training (Utdanningsdirektoratet), which annually develops national tests designed to measure students' core academic skills in numeracy, literacy and English. These tests serve as "unadjusted results" that are only available aggregated at school or municipality level. Although school average results are publicly available, we apply municipality-averages due to the nature of the other variables where the municipal level is the lowest level of data available. Nearly all students participate in these tests, i.e. 98 percent of the students participate in mathematics and English examinations, and 97 percent take part in the reading tests (Udir 2016). The score obtained by each student may be seen as a measure of absolute performance (Sørensen and Geys 2018; De Witte et al. 2014). The main purpose of the tests is to provide educational authorities with information on general student competency, as well as customize teaching in terms of individual needs. In order to ensure individual anonymity, data is missing in some municipalities due to small student populations.

The national tests were first carried out in 2007. This therefore serves as the first year of registry data on school performance in our analysis. In the years between 2007 and 2013, a scale of 1-5 was used to measure the national test score performance. According to the Directorate itself, this method of reporting the test scores was not appropriate for further research. The Ministry of Education therefore implemented a new standard of reporting the tests in 2013, where the scale now ranges from 0-100. Also, the implementation of the tests themselves changed somewhat. The reporting method changed from classical test theory and regular test scores, to item response theory (IRT) with gradual performance levels. The reasoning behind this was that it would be easier to compare a school's performance development over years. Each task in the test would now be attributed a certain performance level. The idea is that one could better describe the student's strengths and weaknesses, make it clear what tasks the student should be able to master, and give the student better feedback with regard to future learning (Fylkesmannen.no 2014). Additionally, it was argued that when using classical test theory, it would be difficult to determine what caused the variations in the test score results. Hence, the technical basis for the national tests changed, so that today IRT-scaling, IRT-linking and equivalences are used from year to year (Udir.no 2016). The registry data is reported aggregated to the municipal level. Our main analysis is based on the average test score results of all three subjects in each municipality.

Due to the different reporting methods, we find it challenging to compare the national test scores from 2007-2013 to the 2014-2015 results. In order to ease the comparison throughout the whole period of 2007-2015, we decided to make a standardization of all test scores when analysing school performance. We therefore estimate a municipality-level percentage deviation from the national average for each year. The standard score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. We do this by subtracting the mean of that year from the observed value, then divide by its standard deviation. The absolute value of what we get represents the deviation between the score and the population mean, in units of the standard deviation. This deviation will then be negative when the score is below the mean, and positive when it is above the mean (Kreyszig 1979). We standardize the average municipal scores based on the following formula:

$$Standardized\ score = \frac{Score - MeanScore}{Standard\ deviation\ of\ score}$$

Test performance indicators

The registry data on school performance indicate that there are great variations from one municipality to another in terms of educational performance. Why these scores differ, however, serves as a source to further discussion. Statistics Norway has conducted a study on students’ performance aiming to examine to which degree schools and municipalities contribute to students’ achievements. This shows that family background such as parents’ education and immigration background have implications for the students' educational skills. The study uses the unadjusted registry data on school performance to estimate indicators for school contribution, where the idea is to control for the composition of students. The municipality-level test performance may be interpreted as the average test score results within a municipality, adjusted for the students' family background. This involves both cross-sectional and value-added (provided for 8th grade only) indicators. The idea is to control for factors that may contribute to school performance that is not determined by the school. The study finds that the indicators show a significantly smaller difference among school performance than the unadjusted registry data imply. Therefore, the observed differences in unadjusted results may be explained by the composition of students (SSB 2017b).

Moreover, the study highlights the importance of taking into account uncertainty when applying the registry data on school performance in further research. We therefore take this into consideration when running our analysis by adding controls for socioeconomic conditions. Furthermore, we adopt the *municipality-level test performance indicator* as the dependent variable in a separate analysis, as this serves as an appropriate source of data that is cleaned for external socioeconomic conditions potentially disrupting the empirical results.

Controls

Although we adjust for within-municipality and yearly effects, there are other explanatory factors that should be controlled for. We argue that the inclusion of these variables strengthens the models.

We furthermore include a set of relevant municipality-level controls when investigating the standardized registry data to counteract potential heterogeneity. The controls include background variables such as (logged) municipality population and the share of students with immigration background. These variables are retrieved from the "Local Government dataset" by Fiva, Halse and Natvik (2017) and by NSD's local government database.

One often sees that research on public policy includes a control for population across entities. There are great variations when it comes to population and size among Norwegian municipalities. The majority of the municipalities have a population about 10.000 while the 100 largest municipalities account for 75 percent of the population (KS.no). In our analysis, we control for population as it might appear to impact the level of teaching, student composition and so on due to the great variation in municipality sizes. The robustness checks will also consider potential outliers in terms of population size. Moreover, large and small municipalities may have unobserved characteristics of their poor/great learning abilities. This is also highlighted in the study by Statistics Norway (2017b).

An assumption is that children coming from an immigration background might face harder obstacles in the academic life compared to non-immigration children. Linguistic challenges, cultural differences, resources at home and so on might be some of the key factors that play into this. That has been a phenomenon for years in many western countries. Many large scale international assessments in recent years have shown that our assumption seems to be true. Program for International Student Assessment (PISA) results indicate that immigrant students often perform at significantly lower levels than non-immigrant students (Hachfeld et al. 2010).

Discoveries have also been made that teachers underestimate how difficult it is for immigration students to overcome the linguistic challenges they are faced with. Hachfeld et. al (2014) indeed find that teachers overestimate the performance of bilingual students, more than the performance of monolingual immigrant or non-immigrant students.

Measurement issues

Throughout the analysis, we must be aware of weaknesses related to potential measurement issues, which should induce carefulness when evaluating the results. One issue is that the Norwegian municipal sector has been subject to continuous structural reforms due to centralization and efficiency improvements. As mentioned, we use the 2012-municipal structure to avoid any issues related to the fact that the number of municipalities changes over the research period. We therefore have N=429 entities throughout the analysis organized as longitudinal data. The time period ranges from 2004 to 2016, and we end up having a total of 5148 observations.

The registry data on school performance is an unadjusted measure of school performance. Standardizing the registry data on school performance makes it possible to compare school performance throughout the whole research period (2004-2015).

Furthermore, we should consider the possible reasons for implementing benchmarking since the variable may be endogenous. That is, some municipalities may implement benchmarking as a result of poor school performance in order to aim for improvements. However, we chose not to accommodate this endogeneity concern due to the limitations of this paper's data, and thereby assume the variables to be exogenous when applying the benchmarking data. How to address such endogeneity concerns in future studies is discussed later on in more detail.

A second element to address is whether the registry data on school performance actually measure school performance, and whether or not it is able to capture the effect of benchmarking. There have been disagreements about the nature of the national tests ever since they were carried out. We do not take part in that discussion, since we believe that is far outside the scope of this thesis. However, we must be aware of the fact that the registry data may not provide the best overview of school performance and its linkage to benchmarking. Due to difficulties finding other appropriate measures on school performance on municipal levels, we apply the registry data with carefulness, as suggested by Statistics Norway. Other related studies also make use of the registry data (see Hovedhaugen et. al. 2017, Sørensen and Geys 2016 among others).

A final element is that the small student masses must be taken into account. Hovedhagen et. al. (2017) study the application of the registry data on school performance in the light of what kind of information one gets of it. They found that only 50 out of 428 of today's municipalities have the required number of students to be able to compare results among themselves and others. That is, in eight out of nine municipalities, the variation in scores were characterized as random, and should thereby be interpreted with carefulness. The study further highlights the improvement of changing the reporting method, as we discussed earlier.

3.3 Descriptive statistics

Table 1 displays the descriptive statistics of the sample. As noted, we use aggregated municipality-level data due to availability. The first section of the table is separated into three time periods of four years each due to the nature of the benchmarking data. The years 2004, 2008 and 2012 represent the years where benchmarking status is reported. We therefore display the period means of test score results in 5th and 8th grade respectively. Hence, the table displays the average test score results for each four-year period as deviations from a municipality mean.

We furthermore convene all the data in the last part of the table. We observe that the standardization of the national test score results generates (means of) test scores centred around zero with standard deviations of (or close to) one. This will be taken into account when analysing the results. The test scores will serve as our main dependent variables. Table A.1 in the appendix displays the development in the national test score results, as well as the municipality-level test performance indicators, over years.

Table 1

Summary statistics

Variable	N	Mean	Std. Dev.	Min	Max
*2004 (2004-2007)**					
Benchmarking (Dummy)	200	0,375	0,4853	0	1
Test Scores, 5th grade**	358	0,0324	1,0004	-3,0148	2,9628
Test Scores, 8th grade**	223	-0,0193	0,9999	-3,7374	3,0579
*2008 (2008-2011)**					
Benchmarking (Dummy)	301	0,2724	0,4459	0	1
Test Scores, 5th grade**	350	-0,0079	1	-5,1886	4,9637
Test Scores, 8th grade**	227	-0,0047	1	-4,792	3,239
*2012 (2012-2015)**					
Benchmarking (Dummy)	330	0,6333	0,4826	0	1
Test Scores, 5th grade**	364	0,013	1	-4,9833	3,9058
Test Scores, 8th grade**	293	0,017	1	-6,9777	5,9902
Total					
Benchmarking (Dummy)	3654	0,46	0,4982	0	1
Test Scores, 5th grade*	3605	0,0047	1	-5,1886	4,9637
Test Scores, 8th grade*	2682	-0,0004	1	-6,9777	5,9902
MLTP-indicator, 5th grade***	2400	3,2962	0,2409	2,4	4,2
MLTP-indicator, 8th grade***	2442	3,4253	0,1618	2,6	4,1
Test Scores, Mathematics, 5th grade	3572	-0,0049	0,248	-2,5526	0,8717
Test Scores, Mathematics, 8th grade	2649	-0,0208	0,2571	-3,1931	0,8728
Student and Municipality characteristics					
Immigration background****	5144	0,0674	0,0408	0,0017	0,3837
(Logged) municipality population	5144	8,4883	1,1502	5,3278	13,3815
Municipality population	5144	11287,69	33898,28	206	647676
Municipalities in the sample	429				

Notes:
The table displays summary statistics of municipality-level data on school performance in the period of 2004-2016, as well as municipality and student characteristics.
* The data in these four-year periods is presented as period-means.
** National test score results on a standardized metric.
*** Municipality-level test performance indicator.
**** Data on students with immigration background is represented as shares of total number of students.

To control for potential confounding factors, we include controls for (logged) municipal population and the share of students with immigration background. Both educational authorities as well as several researchers argue that student masses with a significant share of students with immigration background, might cause variations in school performance (SSB 2017b). The growing share of families with immigration background is distributed highly differently among the municipalities where the bigger cities are often characterized with a larger share of immigration families than rural areas. Figure A.3 in the appendix illustrates the growing number of students with immigration background on a national basis. We observe that the share of students with immigration background increases substantially throughout the sample period. Table A.2 in the appendix displays the control variables in more detail.

Additionally, we perform a separate analysis of numeracy skills as a part of the sensitivity tests. Table A.4 in the appendix displays numeracy scores aggregated at municipal level as yearly averages. These variables are denoted as TS5_Math and TS8_Math*,* and follow similar standardization method as the average test score variables. We include numeracy skills as a separate robustness test as Mathematics are usually associated with significantly greater variations than other subjects due to the nature of learning quantitative skills and adapting to logical thinking (Forskning.no 2015). Also, Mathematical test score results are usually provided with greater media attention due to a general concern that Norwegian students preform poor in numeracy compared to its European comparatives.

Section 4 – Empirical analysis

Our empirical analysis is conducted using a fixed effects method when measuring the effect of benchmarking on educational performance. To the best of our knowledge, there is no similar research on the relationship within the Norwegian municipal sector. However, there is a range of research looking at benchmarking in private organizations that can be related to our study. We also base our empirical approach on recent studies of policy reforms within the public sector, as well as taking advantage of multiple public reports issued by central authorities.

4.1 Fixed Effects

The theoretical analysis implies that introducing benchmarking is likely to be followed by improvements in students' performance. In order to study this relationship, we estimate a fixed effects model where the aim is to control for unobserved confounding factors. We further discuss the idea of controlling for lagged effects due to the component of timing.

We assume that benchmarking status is persistent, meaning that when a municipality implemented benchmarking in one year, it will most likely continue with a benchmarking policy the next year as well. Misreporting benchmarking may serve as a source to measurement error. However, we ignore this possibility due to the fact that the benchmarking data is retrieved from official data sources and is thereby less exposed to misreporting.

Our study focuses on observational variables that vary over time. There will be different municipality effects that may have an impact on national test score results that we are not able to gather. By applying a fixed effects approach, we control for potential omitted variable bias due to variables that are constant over time or across entities. This arises as a feature of using panel data; namely that we can control for all stable characteristics of the entity, i.e. heterogeneity. These characteristics, also referred to as unobserved effects, may be treated as random or fixed effects, depending on whether it is correlated with the explanatory variables or not. When correlation is present, we can apply the fixed effects approach in terms of holding these factors constant (Wooldridge 2002). An example of such factors could be which political party holds the majority, and hence can decide the political agenda. This will be the same for the whole time legislative period, but differ from municipality to municipality. Municipality fixed effects therefore takes care of time-constant unobserved heterogeneity.

Furthermore, we deal with time fixed effects by controlling for variables that are constant across municipalities, but evolve over time. This can refer to policy reforms defined by the central government, for example new education criteria for teachers in schools, updates in the curriculum and so on. We assume that such factors will have the same effect on all municipalities.

"The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics." (Stock and Watson, 2003, p. 289-290)

Baseline models

We let $TS5_{kt}$ and $TS8_{kt}$ denote the national test score results in 5th and 8th grade for municipality *k* at time *t* respectively. Our baseline models, equation (1) and (2), with municipality and time fixed effects are estimated as following:

$$TS5_{kt} = \alpha_0 + \beta BM_{kt} + Controls + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (1)$$

$$TS8_{kt} = \sigma_0 + \beta BM_{kt} + Controls + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (2)$$

Our dependent variables are the standardized test score results in 5th and 8th grade respectively. The parameter of interest, $\beta * BM_{kt}$, is a dummy for benchmarking which is equal to 1 if municipality *k* uses benchmarking at time *t*, and 0 otherwise. Benchmarking is also the main explanatory variable. We further control for population size and students with immigration background as mentioned above.

As noted, the success of these models rest on the assumption that the parameters *k* and *t* hold changes in municipality and year effects fixed over time. Because of this, the models can control for the unobserved heterogeneity as long as this heterogeneity is constant over time. The effect of benchmarking on national test scores can be estimated by treating *k*, the fixed effect, as a parameter to be estimated. The year effect, *t,* is also treated as a parameter to be estimated. The coefficients on dummies for each individual are unobserved individual effects, while the year effects are coefficients on time dummies (Angrist & Pischke, 2008).

When running our fixed regressions, we specify that the standard errors allow for intragroup correlation. This means that we relax the usual requirement that the observations are independent. The observations are still independent across the clusters, but they may not be within these groups. We cluster on municipalities, so this is then the groups to which the observations belong. The reason is that we believe that there could be a correlation across entities, when it comes to the implementation of benchmarking. For example, say a neighbouring municipality begins implementation. One could think that the other municipality notices this, and that it plays a part in their decision of whether or not to also do so. The lowest level of data in our case is Norwegian municipalities, hence, we cluster on them.

The time aspect of implementing benchmarking

Thus far, we have only discussed models that examine the relationship between benchmarking and school performance at the same point in time. However, there is reason to believe that the implementation of benchmarking affects school performance over years since the educational system takes time to respond to such policy changes. This leads us to test whether benchmarking has a *lagged* effect on educational results. There may be a wide range of reasons why some municipalities decide to initiate such policy measures. In order to capture the variation in student performance across municipalities we take heterogeneous effects into account by performing robust regressions.

In many studies, the assumption that the most important omitted variables are time invariant does not seem plausible (Angrist and Pischke 2009). The aspect of time does matter, and thereby needs to be taken into account in our further analysis. When evaluating the implications of benchmarking, we therefore account for the possibility that benchmarking has a delayed effect on educational results. This statistically refers to controlling for earlier performance when running the regression. Moreover, past national test score results might be a time-varying confounding variable that cannot be captured in a time-invariant omitted variable. Students' historical test score results motivate an estimation strategy that controls for the same students' past results directly. Hence, by controlling for earlier student performance as well as testing for benchmarking at an earlier stage, we isolate other confounding factors, and hence find the isolated delayed effect of benchmarking on national test score results. Due to the period of three years in between registry data for 5th and 8th grade, we find it useful to control for benchmarking and student performance three years back in time. The model, equation (3), is estimated as following:

$$TS8_{kt} = \delta_0 + \beta BM_{kt\text{-}3} + TS5_{kt\text{-}3}\ Controls + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (3)$$

Heterogeneous effects

A growing amount of literature has contributed new methods for estimating heterogeneous effects (Grimmer et. al. 2017). This is often related to studies of political policy processes where the aim is to estimate potential treatment effects that vary across sub-populations, i.e. heterogeneous treatment effects. This will in our case correspond to effects of benchmarking that vary across sub-groups of municipalities. We distinguish between small and large populated municipalities when estimating heterogeneous effects.

A unique characteristic of the Norwegian municipal landscape is its diversity when it comes to population size combined with populated rural areas. A core policy within educational policy is to provide primary and lower-secondary education to the whole population regardless of the size of municipality population. As a result, some of the smallest municipalities are characterized with one or very few schools due to small student populations. These municipalities will therefore not be exposed to competition in the same degree as larger municipalities. That is, municipalities with large population will naturally have a greater number of schools within its municipal boundaries. According to the benchmarking theory, this creates a competitive environment among the schools due to the publication and comparison of test score results. This may therefore be reflected in the estimated effect of benchmarking. Hence, we run separate regressions for large (population above 5000) and small (population below 5000) municipalities by applying equation (1) and (2).

Period-average analysis

Due to the set-up of the benchmarking data, we also replicate the analysis by running the models over again with period-average results. That is, since benchmarking is only reported every fourth year, we collapse the registry data on school performance into similar four-year averages.

We collapse our data in three periods. Given that we only have benchmarking data for 2004, 2008 and 2012, we create three time variables for the periods 2004-2007, 2008-2011 and 2012-2015. As noted, our assumption is that a municipality that implements benchmarking will not reverse this in the following years. Hence, we assume that within these periods, this decision will not be reversed, at least until we have for the next period. The data in this case would then be time-period averages. In the case of benchmarking, the variable would then be one if a certain municipality had benchmarking in at least one year within the period.

In this setting, it is even more important to provide estimates that is based on a one-year lag of the benchmarking variable, i.e. the models consider benchmarking in its previous period. The reason follows from a belief that one should not expect benchmarking to cause any effect in test scores in the same period. Therefore, specifying a lag on benchmarking makes it possible to estimate the effect on test scores in the current four-year period of having benchmarking at least some part of the previous four-year period, and implicitly in the entire current period. The estimation strategy follows the same set-up as equation (1) and (2), except that we specify a one-year lag on the benchmarking-dummy.

Municipality-level test performance indicator

In addition to the models presented above, we create a model using municipality-level performance data derived by Statistics Norway. Differences in the composition of students across municipalities are said to have a significant effect on student performance. However, registry data on school performance do not imply any variations in terms of how much the school and the municipality contribute to students' achievements. Statistics Norway therefore conducted a report where such socioeconomic factors are taken into account. Deriving a municipality-level test performance indicator may therefore be interpreted as how the municipality contributes to the students' performance, given that its student mass is average across all characteristics included in the analysis (SSB, 2017b). Hence, the data is adjusted for characteristics such as parents' education level, immigration background, number of students in school, and urban (non-rural) communities. The report highlights that there is correlation between the unadjusted registry data on school performance and the control variables. However, the correlation is removed when testing the municipality-level test performance indicator, and hence, the indicator serves as an appropriate measure on school performance, adjusted for confounding factors. Statistics Norway uses a similar approach when investigating the municipalities' contribution to school performance (SSB, 2017b). We let $MLTP_{kt}$ denote this municipality-level test performance indicator in our models, equation (4) and (5) below, which serves as our alternative dependent variable for school performance. The rest is as before:

$$MLTP5_{kt} = \alpha_0 + \beta BM_{kt} + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (4)$$

$$MLTP8_{kt} = \sigma_0 + \beta BM_{kt} + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (5)$$

As mentioned, the main advantage of applying the municipality-level test performance indicator is that the data is derived by taking confounding factors into account. When applying the municipality-level test performance indicators, we do not consider any controls. We again run fixed effects regressions where the β serves as the parameter of interest.

Critics to fixed effects regression

One side effect of fixed-effects models is that it cannot be used to investigate time-invariant causes of the dependent variables. Technically, time-invariant characteristics of the municipalities are perfectly collinear with the entity dummies. Substantively, fixed-effects models are designed to study the causes of changes within an entity. A time-invariant characteristic cannot cause such a change, because it is constant for each entity.

A common critic towards the fixed effects models is the removal of so called "good variation". This measurement error problem in panel data comes from the fact that the differencing and deviations from mean estimators used to control for fixed effects typically remove both good and bad variation. Put differently, these transformations may remove some of the omitted variable bias, but also remove much of the useful information in the variable of interest (Angrist & Pischke, 2009).

4.2 Results

Based on the theoretical analysis, implementing benchmarking is likely to be followed by improvements in school performance. In this section, we present the estimated regression results. We start off by presenting our baseline models within the fixed effects framework. We further consider lagged effects, heterogeneous effects as well as period-averages, and end the analysis by treating the test performance indictors derived by Statistics Norway as dependent variables.

Baseline regression results

The estimated results suggest that benchmarking, denoted as *BM*, has a very small, almost non-existing effect on test scores for both primary school (5th grade, column 1) and lower secondary school (8th grade, column 2), displayed in table 2. Hence, we clearly see that it would be problematic to conclude that it has any effect at all. The results are insignificant on all levels, i.e. we cannot say with certainty that the coefficients in the model are different from zero. Hence, we cannot say confidently that increased use of benchmarking has an effect on national test scores for 5th graders. For the lower secondary school students, the coefficient is even negative and once more far from statistical significance. However, to initiate a discussion of these differences between both estimates seems a bit unnecessary, given the fact that in both cases we cannot find any significant relationship, and the results may thereby be interpreted as random variations rather than variations due to the implementation of benchmarking.

Table 2 displays the estimated results for the baseline models. Column (1) shows the estimated results of benchmarking on national test score results in primary school (5th grade), while column (2) corresponds to the estimated results for lower secondary school (8th grade). Due to the standardization of the dependent variables, we interpret the results by looking at a one-unit change in *BM*, which is a major event given the (small) standard deviation of the *BM* variable.

Table 2

Baseline regression results

	(1)	(2)
Variables	TS5	TS8
BM	0,0054	-0,116
	(0.068)	(0.071)
Observations	2317	1678
Number of municipalities	378	350
R-squared	0,004	0,015
Control variables	YES	YES
Municipality FE	YES	YES
Year FE	YES	YES
Lagged effects	NO	NO

*Regression results: Baseline models. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

In terms of investigating what happens to the result for a one-unit change in *BM*, the effect size for a one standard deviation change in *BM* is 0,00036 and -0,0082 for the test score results in 5th and 8th grade respectively. That is, these effect sizes are very small compared to the standard deviation of the dependent variables (which are both close to one). Additionally, the confidence intervals vary from -

0,12 to 0,13 (*TS5*) and from -0,25 to 0,02 (*TS8*), implying a satisfying precision of the estimates. However, we cannot conclude whether there is any effect of implementing benchmarking due to insignificant estimates.

Lagged effects

Due to the nature of this study, some student groups are observed twice over the sample period, which allows us to consider lagged effects of benchmarking. That is, 5th graders three years later on will be 8th graders. This means that there is a three-year period where we can analyse how benchmarking has changed the scores from when they were three years younger. Say, a municipality implements benchmarking in 2007. Their 5th graders will then be 8th graders tested in 2010. This is something that the baseline models do not capture. Hence, we estimate another model that takes this three-year delayed effect into account. The specification of the control variables is as before, while we also include the national test score results for 8th graders at time *t,* obtained as 5th graders at time *t-3*. The regression result of the lagged effects model is displayed in table 3.

Table 3

Lagged effects regression results

	(1)
Variable	TS8
L3.BM	-0,0387
	(0.073)
L3.TS5	0.4384***
	(0.053)
Observations	1052
Number of municipalities	318
R-squared	0,242
Control variables	YES
Municipality FE	YES
Year FE	YES
Lagged effects	YES

*Regression results: Lagged effects models. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

The results are much the same as the ones we get from our baseline fixed regression for the 8th graders. The benchmarking coefficient is still negative, but now moves even closer to zero. Clearly we see that there is a significant connection (at the 1 percent level) between 8th grade scores and the 5th grade scores from three years back, which seems reasonable. It helps since it adds explanatory power to our model. However, the lagged effects model confirms the baseline results in that all our models seem to point towards the fact that we cannot reject the null hypothesis that the coefficients are different than zero. In other words, given our models, it seems that we cannot confidently claim that benchmarking is associated with better school performance. The narrow confidence intervals (-0.18 to 0.10) tell us at that with a high degree of certainty we can say that the effect is nearly zero in all three cases, i.e. the observed effect may be random and not linked to the implementation of benchmarking. Looking at a one-unit change in the *BM*, the effect size of a one standard deviation change in *BM* is -0.0028 for the estimated lagged effect for *TS8*. Again, compared to the standard deviation of *TS8*, this is a very small effect.

Heterogeneous effects

Taking heterogeneous treatment effects into consideration in terms of running separate regressions for small and large sub-groups of municipalities add to the overall findings more or less by confirming that the introduction of benchmarking does not imply improved school performance. Within the sub-group of small municipalities, the estimate of benchmarking is statistically significant (on 10 percent level) and slightly negative for test scores obtained in 8th grade (column 2), illustrated in table 4. Changing educational policy, i.e. implementing benchmarking will in that case change the overall results negatively by 0,15 on average. The estimates for 5th grade test scores do not imply any significant changes in overall test score results (column 1).

Municipalities with a population above 5000 are exposed to a greater level of competition within the municipalities, which may affect the benchmarking estimates. The less negative estimate for test scores obtained in 8th grade (column 4) may be explained by the increased degree of competition. Also, the estimates for 5th grade (column 3) add to the analysis by confirming that benchmarking does not imply any changes in test score results. Hence, heterogeneous effects may be related to less negative test scores obtained in 8th grade due to increased competition on school performance within subgroups of larger municipalities.

Table 4

Heterogeneous treatment effects

	Small population		*Large population*	
	(1)	(2)	(3)	(4)
Variables	TS5	TS8	TS5	TS8
BM	0,0045	-0.1520**	0,0081	-0,0371
	(0.079)	(0.055)	(0.114)	(0.179)
Observations	1270	1060	1048	618
Number of municipalities	194	186	194	169
R-squared	0,008	0,021	0,007	0,019
Control variables	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Lagged effects	NO	NO	NO	NO

*Regression results: Heterogeneous effects. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

Period-average analysis

The period-average analysis, somewhat surprisingly, imply opposite results compared to our previous regression results. Both student groups are here characterized by negative coefficients. The estimate for 5th grade students is even significant at the 10 percent level. Possible reasons for these surprising results may be related to the fact that there now are fewer observations of benchmarking in the sample due to only being observed over three periods for each municipality. One may also add to the fact that there is a large amount of missing observations in the benchmarking data, which may explain the variety in the estimated coefficients. The regression estimates are displayed in table 5.

Table 5

Period-average regression results

Variable	(1) TS5	(2) TS8
L.BM	-0.2752**	-0,0248
	(0.089)	(0.082)
Observations	465	412
Number of municipalities	329	315
R-squared	0,078	0,047
Control variables	YES	YES
Municipality FE	YES	YES
Year FE	YES	YES
Lagged effects	YES	YES

*Regression results: Period-average data. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

4.3 Extended analysis

As noted, we apply the municipality level test performance indicator as an extended part of our analysis. The municipality-level test performance may be interpreted as the average test score results within a municipality, adjusted for the students' earlier results as well as family background (SSB 2017b). We treat these indicators as the dependent variables in another fixed effects regression. Since the data contains much of the information that explains school performance, we exclude our initial control variables. We therefore end up analysing the effect of benchmarking on the municipalities' contribution towards schooling. The regression results are displayed in table 6.

The findings are consistent with what we obtained in our previous models. For both of the student groups, the effect is close to zero, and marginally negative for 8th graders. Although the regression estimates are still insignificant at all levels, the results add to our analysis since it strengthens the overall insight of the noneffect of benchmarking. Also note that the signs appear consistent throughout the analysis.

Table 6

Extended analysis results

	(1)	(2)
Variable	MLTP5	MLTP8
BM	0,0168	-0,0321
	(0.026)	(0.020)
Observations	1822	1832
Number of municipalities	374	373
R-squared	0,018	0,012
Control variables	NO	NO
Municipality FE	YES	YES
Year FE	YES	YES
Lagged effects	NO	NO

*Regression results: Extended analysis. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

Determining whether there are sizable effects, we again look at the result of a one-unit change in *BM*. The extended analysis obtains effect sizes for a one standard deviation change in *BM* of 0.0004 (*MLTP5*) and -0.0006 (*MLTP8*), which again are minuscule compared to the standard deviation of the dependent variables. The fractions add to the analysis by showing a somehow consistent effect of zero on 5th graders, while the effect on 8th graders remains slightly negative. In other words, regardless of the insignificant coefficient estimates, we observe that 8th graders may be affected slightly negative by the implementation of benchmarking, although this is effect is minor.

4.4 Robustness

We expose the regression results above to several robustness checks to test the validity of the analysis. First, in addition to investigating the overall student performance by standardizing the municipal average of educational results, we examine the effect of benchmarking on results in Mathematics alone. We continue by excluding municipalities with larger and small populations as these might serve as outliers in terms of population size. Finally, we include a study of the within period-average analysis, where we treat the benchmarking variable based on assumptions due to the large amount of missing data. This set of various models serves as a measure on sensitivity and precision of the initial analysis.

Isolated analysis of numeracy skills

Mathematics is usually reported with significantly greater variations than reading and English abilities. A range of research points out that children adopt to mathematical abilities differently due to the nature of learning quantitative skills. Mathematics is based on arguments, evidence and generalization and follows from strict rules. One may find it hard to compare and relate to real life events (Forskning.no 2015). The model, equation (6) and (7), is specified with Mathematics-results as the dependent variable while the rest is as before.

$$TS5_{Mathematics} = \alpha_0 + \beta BM_{kt} + Controls + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (6)$$

$$TS8_{Mathematics} = \sigma_0 + \beta BM_{kt} + Controls + \lambda_k + \gamma_t + \varepsilon_{kt} \qquad (7)$$

The fixed effects regression estimates show negative coefficients for both student groups. For the 5th graders, we (again) get a coefficient as low as basically zero, but with negative sign. Again, it seems that we can confidently claim that the coefficient is not different from zero, hence, there is no significant effect of benchmarking on 5th grade Mathematics.

Table 7 displays the regression results for numeracy skills alone. We observe that we now obtain statistically significant estimates for the 8th graders at the 5 percent level (column 2). That is, the estimate confirms that the coefficient is marginally negative. Hence, implementing benchmarking influences 8th grade math scores slightly negative in most cases. In this case, the effect size of a one-unit change in *BM* is -0.00002 (*TS5_Math*, column 1) and -0.0015 (*TS8_Math*, column 2), i.e. very small effect. Also, the confidence intervals continue being narrow with variations of -0.03 and 0.03 for *TS5_Math* and -0.1 and 0.004 for *TS8_Math*, implying precise estimates.

Table 7

Regression results for Mathematics

	(1)	(2)	(3)
Variable	TS5_Math	TS8_Math	TS8_Math
BM	-0,001	-0.0561*	0,028
	(0.016)	(0.026)	(0.028)
Observations	2293	1655	1026
Number of municipalities	379	355	313
R-squared	0,004	0,044	0,102
Control variables	YES	YES	YES
Municipality FE	YES	YES	YES
Year FE	YES	YES	YES
Lagged effects	NO	NO	YES

*Regression results: Extended analysis. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

However, the results slightly change when we control for the lagged effects of benchmarking (column 3). Just as in the case with the average test score results, we take into consideration the math scores that the 8th graders obtained as 5th graders in time *t-3*. When doing so, the results change. Instead of a significant negative coefficient, we now get an insignificant effect close to zero. This is consistent with what we obtained in our baseline model. Although, in that case the effect is always statistically insignificant.

Exclusion of outlier-municipalities

Second, we re-estimate all models by excluding municipalities with a population size below 400 and above 90.000 inhabitants. The municipalities excluded serve as obvious outliers in terms of population size in the Norwegian municipal landscape. This robustness strategy is adopted by Sørensen and Geys (2016). Excluding the following municipalities Oslo, Bærum, Bergen, Trondheim, Stavanger, Utsira and Modalen does not imply any changes in our estimates. The results largely confirm the (non-existing) effect in the regression analysis. A visible comparison of the estimates is provided in table A.5 in appendix.

The estimates imply that dropping outlier municipalities due to either small or large population sizes does not cause any visible difference in our estimates. Our results are therefore not driven by the student performance in the few very small or large populated municipalities.

Period-average analysis

Due to the collapsing of data, we get many missing observations with regard to the benchmarking dummy. As a robustness check, we introduce the following assumption to retrieve some of these missing values: We assume that a municipality only changes from not having benchmarking to implementing it, once. In other words, they do not go back once they have decided to implement it. We further specify a one-period lag on the benchmarking dummy. As noted, the one-period lag effectively estimates the effect on test scores in the current four-year period when a municipality has had benchmarking at least some part of the previous four-year period. The fixed effects regression results of this somewhat experimental method is displayed in table 8 below.

We observe that this 'rescuing' some of the observations in this way give us significant (at the 10 percent level) negative estimate for 5th grade test scores (column 1). Recall that the (less experimental) period-average analysis above also implied a similar relationship. We again observe an insignificant estimate for 8th grade performance. These findings confirm, on the whole, that all results must be treated with carefulness. We can therefore reject our initial hypothesis that the impact of benchmarking is positive since the estimates usually turns out to be zero, and often even slightly negative.

Table 8

Experimental period-average regression results

Variable	(1) TS5	(2) TS8
L.BM	-0.2700**	0,0187
	(0.098)	(0.101)
Observations	615	517
R-squared	0,004	0,022
Number of municipalities	356	346
Control variables	YES	YES
Municipality FE	YES	YES
Year FE	YES	YES
Lagged effects	YES	YES

*Regression results: Experimental period-average analysis. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*

Finally, one may argue that the set of our estimated models serves as a sensitivity analysis by itself due to the different regression and data set-ups. By estimating school performance relative to the impact of benchmarking using a range of different variables and measurements on educational results, the (more or less) consistent results make up the validity of the analysis.

Section 5 - Discussion

This thesis is an analysis of the impact of benchmarking on educational performance in primary- and lower secondary schools in Norwegian municipalities. Our findings indicate that implementing benchmarking does not link to significant improvements of educational results, as initially suggested by our theoretical analysis. Thus, we are not able to confirm our main hypothesis that benchmarking serves as a source to improved educational performance.

In terms of investigating potential factors that are said to have an impact on educational performance, we considered the share of students with immigration background. This is highlighted as one of the key factors to impact how children are performing at school by a wide range of studies (Statistics Norway among others). Future research may also consider controlling for the level of higher education among the students' parents. This is said to have an impact on how children perform at school (SSB 2017b).

When it comes to the municipality level test performance indicator as a part of our extended analysis, these variables also take into account immigration background in addition to a range of other factors that are said to have an impact on school performance. Although such characteristics explain the variety in school performance, we were not able to link it up to the impact of benchmarking. Hence, the extended analysis adds to the overall findings by reporting similar results as the baseline models, and thereby not confirm our hypothesis.

A key lesson from our analysis is that when taking into account lagged effects, controlling for 5th grade scores clearly has a high explanatory power for 8th graders' performance. A similar relation is found when investigating numeracy skills separately. We see that for both the overall results, as well the separate analysis on Mathematics, there seems to be a slight negative effect of benchmarking, although not significant. This changes when we control for the lagged effect of benchmarking. A potential explanation could be that skilled students are not affected positively by benchmarking, and thereby counteract the potential positive effect of poor performing students. This somehow smooths out the effect of benchmarking, and we are left with an effect of close to zero although there might be changes across the different student types that we are not able to capture. This may also be the case in other parts of the analysis. However, analysing the effect of benchmarking across sub-groups of students clustered by academic scores requires confidential, individual-level data.

5.1 Limitations

We have only assumed that implementing benchmarking affects school performance. However, there may be a reversed relationship in terms of that some municipalities implement benchmarking as a result of poor educational performance. This may cause endogeneity problems in the analysis that needs to be taken care of. Although we limit this thesis by not accommodating potential endogeneity concerns any further, we suggest that future analysis should assess this problem. One way to deal with such endogeneity is to conduct an instrumental variable (IV) estimation where the aim is to find appropriate instruments that fulfil the two key assumptions. The central assumption underlying the validity of an IV analysis rests on that the instrument fulfils the exclusion restriction of only affecting the outcome of interest, as well as being a strong predictor of the relevant explanatory variable, i.e. be a sufficiently strong instrument (Sørensen and Geys, 2016). However, one needs to be aware of the difficulty of finding an instrument that excludes any direct influence of school performance. Sørensen and Geys (2016) accommodated endogeneity problems in their study by using both hydropower income as well as an alternative set of geographical instruments when explaining the relationship of Norwegian municipal revenues and outsourcing decisions.

Additionally, one may identify whether there is a pattern of which political coalitions that holds the power and thereby makes the decision of implementing performance measures such as benchmarking. One hypothesis is that conservative-liberal local governments are more willing to implement public available performance measures. This may also be the fact within the teacher unions that may be negative to policy reforms and structural changes. This may contribute to the analyses of the impact of performance information in public organizations by Sørensen and Geys (2018) among others.

Future studies within this field should consider looking into the effect of benchmarking on the individual student level. The data we were able to obtain was based on educational results from all the Norwegian municipalities. That is, public available data which is relatively easy to access. If one were able to see how increased focus on performance comparisons within public education affects each student individually, it would certainly add valuable information to this field of research. However, such a study would require personal data, which is not as accessible and would require a more in-depth approach. This would, however, enable such educational policy studies to identify student types, as well as take advantage of those sub-groups throughout the analysis.

Section 6 - Conclusion

This analysis contributes to the recent literature on public policy reforms, in particular the educational field. In this thesis, we evaluated whether introducing benchmarking induces improvements of educational results in primary and lower-secondary schools. Previous studies indicate that such performance measurements imply improved results (Gerrish 2016, Greve 2007). Given the fact that Norwegian municipalities to an increasing extent have implemented benchmarking, this remains an important question also from a policy perspective. Our main findings indicate that public available measurements of performance are not necessarily associated with improved results, and hence, we fail to confirm our hypothesis.

Clearly, our analysis is specific to the Norwegian setting, and hence may raise concerns about the general nature of the results and thereby the implication of it. The findings in this thesis does not only contribute to the existing literature on policy reforms at the local level in public organizations, but also add to the discussion of the educational sector in terms of measuring school performance and quality assessment. As we fail to find any significant relationship linking benchmarking and school performance, our results suggest that measurement performance decisions are partly irrelevant for the following results. To improve our understanding of what drives public policy processes in terms of which municipalities (and why) decide to introduce measurement performance methods, the existence of gradual dynamic developments in the institutional framework, public authorities' service provision serve as an important field for future research (Sørensen and Geys 2016).

Finally, studying the relationship behind the variations in school performance allowed us to take part in a complex discussion. As long as the enrolment in public schools remain at such substantial level, this analysis adds to the current processes of restructuring and efficiency initiatives. Future empirical studies on performance measurements of educational results would be very valuable in terms of gaining a deeper understanding of our educational system, and how it responds to policy changes.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometric: An empiricist's companion*. New Jersey: Princeton University Press.

Baekgaard, Martin., Serritzlew, Søren. (2016). "Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension." *Public Administration Review*, Vol. 76, Issue 1, p 73-82

Borge, L.-E., Falch, T. and Tovmo, P. (2008). Public Sector Efficiency: The Roles of Political and Budgetary Institutions, Fiscal Capacity, and Democratic Participation. *Public Choice* 136, 475-495.

Boyne et al. (2009). Democracy and Government Performance: Holding Incumbents Accountable in English Local Governments. *Journal of Politics*, Vol. 71, Issue 4, p 1273-1284

Business Dictionary (2017). Benchmarking. Retrieved from: http://www.businessdictionary.com/definition/benchmarking.html

Cameron, A. Colin; Trivedi, Pravin K. (2005). Microeconometrics: Methods and Applications. Cambridge University Press. Pp. 717-19.

De Witte, Kristof., Geys, B. and Solondz, C. (2014) Public Expenditures from a Policy Intervention in the Netherlands. *Economics of Education Review* 40: 152-166.

Figo, David & Loeb, Susanna. (2011). School Accountability. Handbook of the Economics of Education, Volume 3

Fiva, Jon H., Askill H. Halse and Gisle J. Natvik (2017): Local Government Dataset. Available at www.jon.fiva.no/data.htm

Fylkesmannen.no (2014). "Ending i nasjonale prøver – Ny skala og måling av utvikling over tid". Retrieved from: https://www.fylkesmannen.no/Documents/Dokument%20FMOA/Barnehager%20og%20opplæring/Grunnskole%20og%20videregående%20opplæring/Endringer%20i%20nasjonale%20prøver_FMOA%20250914.pdf

Gerrish, Ed. (2016). The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis. *Public Administration Review,* Vol. 76, Issue 1, p 48-66

Greve, Henrich. (1998). Performance, Aspirations and Risky Organizational Change. *Administrative Science Quarterly,* Vol. 43, No. 1 (Mar., 1998), pp. 58-86

Greve, Henrich (2003). *Organizational learning from performance feedback: A behavioural perspective on innovation and change*. Cambridge University Press

Greve, Henrich. (2007). Hvordan lærer organisasjoner av resultatmåling? *MAGMA*

Grimmer, J., Messing, S., & Westwood, S. (2017). Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. *Political Analysis,* 25(4), 413-434

Grønn, E. (2008). *Anvendt mikroøkonomi.* Oslo: Cappelen akademisk forlag.

Hachfeld et al. (2010). Does immigration background matter? How teachers' predictions of students' performance relate to student background. *International Journal of Educational Research*, Vol. 49, issues 2-3, pp 78-91.

Hovedhaugen et al. (2017). National test results: representation and misrepresentation. Challenges for municipal and local school administration in Norway. *Nordic Journal of Studies in Educational Policy,* Vol. 3, issue 1

Iris BenDavid-Hadar og Adrian Ziderman (2011): A New Model for Equitable and Efficient Resource Allocation to Schools: The Israeli case. *Education Economics,* 19:4, side 341-362

James, Oliver., John, Peter. (2007). "Public Management at the Ballot Box: Performance Information and Electoral Support for Incumbent English Local Governments. *Journal of Public Administration Research and Theory,* Vol.17, Issue 4, p 567-580

James, Oliver., Moseley, Alice. (2014). Does Performance Information about Public Services Affect Citizens' Perceptions, Satisfaction and Voice Behaviour? Field Experiments with Absolute and Relative Performance Information. *Public Administration,* Vol. 92, Issue 2, p 493-511

Kommunesektorens Organisasjon (2016): Dette må du vite om kommunereformen. Retrieved from: http://www.ks.no/fagomrader/samfunn-og-demokrati/kommunereformen/fakta-om-reformene/dette-ma-du-vite-om-kommunereformen/

Kunnskapsdepartementet. (2017) "Rettferdig og effektiv finansiering av grunnskolen" Retrieved from https://www.regjeringen.no/no/dokument/dep/kd/rapporter_planer/aktuelle-analyser/aktuelle-analyser-om-andre-tema/rettferdig-og-effektiv-finansiering-av-s/id661112/

Kreyszig, E. (1979) *Advanced Engineering Mathematics.* New York: Wiley

Larsen, Per Kristian (2013). Endring i nasjonale prøver – ny skala og måling av utvikling over tid. Retrieved from https://www.fylkesmannen.no/Documents/Dokument%20FMOA/Barnehager%20og%20opplæring/Grunnskole%20og%20videregående%20opplæring/Endringer%20i%20nasjonale%20prøver_FMOA%20250914.pdf

Leuven, Edvin., Hessel Oosterbeek and Marte Rønning (2008), Quasi-Experimental estimates of the effect of class size on achievement in Norway. Norwegian University of Science and Technology. No 2/2008. Retrieved from http://www.svt.ntnu.no/iso/wp/wp.htm

Ministry of Knowledge and Education (2011). ”Rettferdig og effektiv finansiering av skolen”. Retrieved from https://www.regjeringen.no/no/dokument/dep/kd/rapporter_planer/aktuelle-analyser/aktuelle-analyser-om-andre-tema/rettferdig-og-effektiv-finansiering-av-s/id661112/

Moynihan, Donald P. (2008). The Dynamics of Performance Management: Constructing Information and Refom. Washington DC: *Georgetown University Press*

Nielsen, Poul., Baekgaard, Martin. (2015). Performance Information, Blame Avoidance, and Politicians' Attitudes to Spending and Reform: Evidence from an Experiment. *Journal of Public Administration Research and Theory.* Vol. 25, Issue 2, p 545-569

Norsk institutt for by - og regionforskning (2004). Kommunal organisering 2004: Redegjørelse for kommunal- og regionaldepartementet database. Retrieved from https://www.regjeringen.no/no/dokumenter/kommunal-organisering-2004/id106064/

Norsk institutt for by - og regionforskning (2008). Kommunal organisering 2008: Redegjørelse for kommunal- og regionaldepartementet database. Retrieved from https://www.regjeringen.no/globalassets/upload/KRD/Rapporter/Rapporter_2012/2012-21.pdf

Norsk institutt for by - og regionforskning (2012). Kommunal organisering 2012: Redegjørelse for kommunal- og regionaldepartementet database. Retrieved From https://www.regjeringen.no/globalassets/upload/krd/vedlegg/komm/kommunal_organisering_nibr.pdf?id=2123020

Norsk institutt for by - og regionforskning (2016). Kommunal organisering 2016: Redegjørelse for kommunal- og regionaldepartementet database. Retrieved from https://www.regjeringen.no/contentassets/30c1810758ab462581d00fbd7ec75425/kommunal_organisering_2016.pdf

Nøra, Stig (2015) “Hvorfor er det så vanskelig med matte?” Retrieved from https://forskning.no/skole-og-utdanning/2015/09/hvorfor-er-det-sa-vanskelig-med-matte

Propper, Carol, & Deborah Wilson (2003). The use and usefulness of performance measures in the public sector. *Oxford Review of Economic Policy,* Vol. 19, No. 2

Ramsey, F., Schafer, D., (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd ed.* Duxbury Press

Roald, K. (2010). *Kvalitetsvurdering som organisasjonslæring mellom skole og skoleeigar.* Doktoravhandling, Universitetet i Bergen.

Skoleporten. Nasjonale prøver barne- og ungdomstrinn (2017). Retrieved from https://skoleporten.udir.no/rapportvisning/grunnskole/laeringsresultater/nasjonale-proever-ungdomstrinn/hedmark-fylke?enhetsid=04&vurderingsomrade=11&skoletype=0&utdanningstype=--&skoletypemenuid=0&underomrade=51&sammenstilling=12&fordeling=2&enhetsfilterid=8c08bbcb-2a52-4626-bc94-7c5913123b5f

Statistisk sentralbyrå (2016). Sysselsatte i kommunal sector 2016, 4. kvartal. Retrieved from https://www.ssb.no/arbeid-og-lonn/statistikker/komregsys/aar/2017-03-15

Statistisk sentralbyrå (2017a). Elevar i grunnskolen. Retrieved from https://www.ssb.no/utdanning/statistikker/utgrs

Statistisk Sentralbyrå (2017b). Hvor mye bidrar skolen til elevenes læring? retrieved from https://www.ssb.no/utdanning/artikler-og-publikasjoner/hvor-mye-bidrar-skoler-til-elevers-laering

Stock, James H., and Mark W. Watson (2012). *Introduction to Econometrics. 3rd ed.* The Pearson series in economics: Boston, Mass.: Pearson.

Sørensen, Rune J. & Benny Geys (2016). Revenue Scarcity and Government Outsourcing: Empirical Evidence from Norwegian Local Governments. *Public Administration,* Vol 94, Issue 3, Pages 577-861.

Sørensen, Rune J. & Benny Geys (2018). Never change a winning policy? Public Sector Performance and Politicians’ Preferences for Reforms. *Public Administration Review, In Press.*

Utdanningsdirektoratet (2016) Metodegrunnlag for nasjonale prøver. Retrieved From https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover.pdf

Utdanningsdirektoratet. (2017). Rammeverk for nasjonale prøver. Retrieved from https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/

Zachrisen, Oda Opdal and Steffensen, Kjartan (2016). Dokumentasjonsnotat om skole- og kommunebidragsindikatorer i grunnskolen. Retrieved from:https://www.ssb.no/utdanning/artikler-og-publikasjoner/_attachment/286980?_ts=158d896b040

Appendix

Table A.1

Year	TS5	TS8	SSB_ML5	SSB_ML8
2004	.	.	.	.
2005	.	.	.	.
2006	.	.	.	.
2007	0,0324634	-0,0193193	.	.
2008	-0,0026022	0,0147101	.	.
2009	-0,0209964	-0,0116713	.	.
2010	0,0263752	-0,0006343	3,323399	3,433252
2011	-0,0109598	-0,0211368	3,323399	3,433252
2012	0,0169326	0,0108938	3,292982	3,427251
2013	0,0083534	0,0241299	3,292982	3,427251
2014	0,0000042	0,0000234	3,271646	3,415212
2015	0,0000032	0,0003022	3,271646	3,415212
All (mean)	0,00543069	-0,00076908	3,296009	3,425238333
All (Std. Dev)	0,016334127	0,014502863	0,023263109	0,008216983

The table describes standardized, yearly average test score results (TS5 and TS8), as well as the municipality-level test performance indicators derived by Statistics Norway.

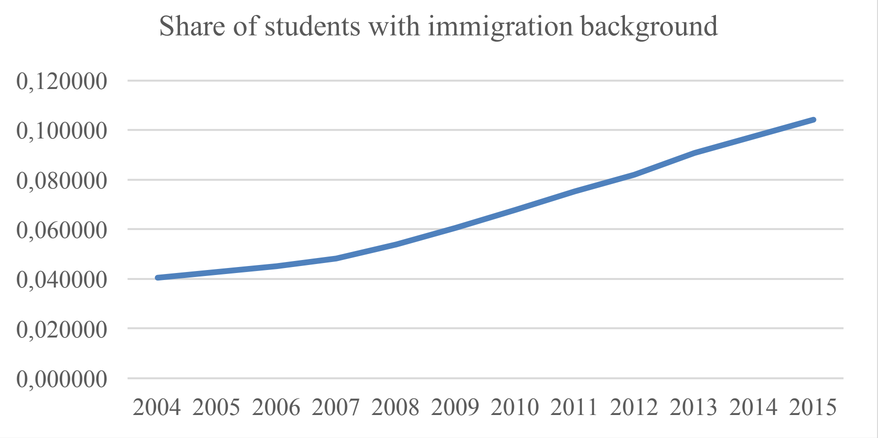
Table A.2

Control variables: Municipality and student characteristics

Year	log_Population	Immigration background
2004	8,473936	0,040599
2005	8,472913	0,0428543
2006	8,472255	0,0451184
2007	8,471871	0,048131
2008	8,475361	0,0538359
2009	8,479403	0,060516
2010	8,485021	0,0679082
2011	8,491159	0,0753327
2012	8,493752	0,0821143
2013	8,51017	0,0908676
2014	8,51479	0,097612
2015	8,518474	0,1042108
All (Mean)	8,488334	0,0674047
All (Std.dev)	1,150221	0,0408122

The table shows the yearly levels of (logged) population and share of students with immigration background. Both variables are municipal averages, and will serve as control variables.

Figure A.3



The figure illustrates the development in students with immigration background over the sample period.

Table A.4

National test scores in Mathematics

Year	TS5_Math	TS8_Math
2004	.	.
2005	.	.
2006	.	.
2007	-0,0086452	-0,0438845
2008	-0,0184334	-0,0698314
2009		-0,01632
2010	-0,0009551	-0,0869605
2011	-0,0211688	-0,020659
2012	-0,0013102	-0,0008896
2013	-0,0015271	-0,0100093
2014	0,0004343	-0,0000403
2015	0,0003022	0,000413
All (mean)	-0,005732822	-0,02479777
All (Std. Dev)	0,008457147	0,03163935

The table displays the standardized, average national test score results obtained in Mathematics each year.

Table A.5

Robustness check: Removing outlier municipalities

	(1)	(2)	(3)	(4)	(5)
Variables	TS5	TS8	TS8	MLTP5	MLTP8
BM	0,0034	-0,1137	-0,0542	0,0188	-0,0318
	(0.069)	(0.073)	(0.075)	(0.026)	(0.020)
Observations	2271	1633	1022	1792	1802
Number of municipalities	372	345	313	369	368
R-squared	0,004	0,015	0,243	0,02	0,012
Control variables	YES	YES	YES	NO	NO
Municipality FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
Lagged effects	NO	NO	YES	NO	NO

*This table reports the regression results obtained in the robustness checks after removing outlier municipalities. Robust standard errors clustered on municipalities in parentheses. ***p<0.01, **p<0.05, *p<0.1*