Preliminary Master Thesis Report at

BI Norwegian Business School

- Does Benchmarking Have a Positive Effect on Educational Results? -

Hand-in date:
15.01.2018

Examination code and name:
GRA 19502 Master Thesis

Name of supervisor:
Rune Sørensen

Programme:
Master of Science in Business – Major in Economics

Content

Summary

The recent years have seen an increasing trend when it comes to measuring the level of performance in the public sector. This practice may be seen as a result of common practice in the private sector, where measurement methods are frequently used to determine organizational performance. One way of doing so is to compare own performance to historical results, which we often refer to as *Benchmarking*. As benchmarking now also serve as a source of measuring performance in the public sector, we decided on investigating the usage of benchmarking on educational performance on the municipal level.

By using evidence from the private sector together with recent research from the public sector, we want to analyse whether benchmarking has had a positive effect on national test score results. We have collected data from multiple public available resources, which we use as a base for our models. Based on the nature of fixed effect regression, we aim to analyse the effect of benchmarking while holding other factors fixed. We therefore compare municipalities with themselves before and after benchmarking was implemented. Our main hypothesis is that benchmarking has had a positive effect. We base our beliefs from findings in private organizations as well as recent research on public policy decisions.

Introduction

Norwegian primary schools use the so-called “national test score results” to determine the students’ level of knowledge in mathematics, English and reading in 5th and 8th grade. These test scores are used in order to evaluate quality and development at several levels in the Norwegian school system. Further, national test score results are used as a comparison at the national level to investigate and identify possible quality differences among students, schools and municipalities.

Measurement methods of performance are frequently used within the private sector to determine organizational performance based on comparisons with historical results of themselves. This method of comparison is often referred to as *Benchmarking.* Although benchmarking is less used within the public sector services to determine performance, recent practice show an increasing trend in this regard. Therefore, our belief is that the implementation of benchmarking as a measurement of performance in Norwegian primary schools provide a positive effect on national test score results.

Motivation and Research Question

Our research question is as follows:

- *Does Benchmarking have a positive effect on educational results?*

The motivation of this paper is to look at the effect of benchmarking on national test scores for students in the 5th and 8th grade in Norwegian public schools. Nationals tests are used to determine the level of reading, mathematics and English abilities for these students.

Using business related definitions; benchmarking is seen as measuring the quality of an organization's programs, products, policies, strategies, and then comparing them to standard measurements or similar measurements of its piers. The objective in many cases is to determine what and where improvements are called for, to analyse how other organizations achieve their performance levels, and to use information to improve their performance levels (Businessdictionary.com, 2017).

Performance evaluations and benchmarking are something that have been used extensively in businesses and corporations throughout the years. It is used as a motivational factor to raise performances, should they drop below a requested level. It is a topic that has been covered both in business administration theory as well as in practice. We will use this in the context of public sector performance. Rather than looking at the impact of benchmarking on corporate performance, we will analyse the impact on public administration performance. This in turn can determine performance levels on the national tests.

In recent years there has been an increase in the number of municipalities that have chosen to use benchmarking to supervise their performance on provision of public goods (NIBR, 2016). We will look at how this has affected the national test scores. The goal is to see whether there is a statistically significant effect between benchmarking and national test scores. To a large extent, it is fair to assume that

the goal of the public sector is to implement an educational policy that maximizes the performance of the students. In this context, that is reported through the national test scores. With regard to the public sector there has not been as much research when it comes to benchmarking. In a human capital perspective, the findings in this paper will also add to the discussion of educational effects on future labour market outcomes. If benchmarking has a significant effect on national test scores, it could mean that the implementation of it can improve the human capital stock of the future.

Hypotheses

Main hypothesis

- *Benchmarking has a positive effect on educational results.*

Our main hypothesis is based on that we find it reasonable to believe that implementing benchmarking has a positive effect on national test score results. We base our beliefs on evidence from the private sector, which we discuss below, expectations about competitiveness and recent experimental work on responsiveness to performance information in the public sector. Being able to compare test score results provide a public available overview of school performance on municipal levels. Reaching the benchmark will in our case serve as a level of satisfaction, given historical results. Additionally, the component of competition among municipalities may also be a source of motivation.

Recent experimental work on responsiveness to performance information has shown that positive information may serve as a level of "satisfaction". On the other hand, negative information on performance tends to have to a greater impact on the degree of responsibility by superiors, in particular elected officials (Sørensen & Geys, 2017: Nielsen and Moynihan 2017). As we are investigating how the increasing usage of benchmarking may explain improvements in national test score results, we aim to examine this effect, which preferably supports our main hypothesis.

One may argue that performance and governance through specific measurements apply to the private sector only due to targeting profits. However, there has gradually been common to refer to benchmarking in the exercises provided by the public sector in order to make comparisons about quality as well as a source of improvement. As described above, we aim to look at the use of benchmarking by municipalities, and how it may has improved national test scores in public schools. Based on positive findings from the private sector, our hypothesis says that benchmarking has had a positive effect on national test score results.

Another factor to include when discussing the use of benchmark is the position of (public) schools in the society. One may argue that Norwegian primary schools hold monopoly power in terms of providing education. There is little competition in the education field, which may impact the quality as well as the ability to improvements. Further, little degree of competition may also indicate that the level of reaching a satisfactory test score is too unambitious. Additionally, the teachers' unions hold a strong position, which again may serve as a restriction to restructuring. Therefore, by introducing benchmarking as an instrument of comparison between municipalities may serve as a good model, as well as a motivation to improvement.

The Norwegian upper secondary school is concerned with relatively high level of dropouts (27% in 2017, ssb.no 2017). This challenge is to be addressed by a number of tools in order to reduce the number of dropouts. On an individual level, national test score results in 8th grade play an important role when it comes to identifying less skilled students and students with special needs. Earlier studies have also touched upon the relationship between test score results in 8th grade and upper secondary school dropouts. Researchers have questioned the fact that a large share of 8th graders have lacking skills in reading particularly, which is expected to be a fundamental skill by the age of an 8th grader. When addressing the problem of upper-secondary school drop-outs, national test score results are used as a base of action towards this issue (Statsministerens kontor, 2009). There is reason to believe that students with poor results in lower secondary school have higher probability of dropping out in upper secondary school (Nordlys.no, 2010). If one already lack behind academically at an early stage, this might be difficult to

catch up with when the academic level is higher. When it comes to supporting research on such beliefs, both SINTEF and several research institutes have at an earlier point identified this group as the "critical-group" of potential future dropouts (Udir.no, 2007, utdanningsnytt.no, 2015).

One key component in this case is how this share of "lost" future human capital affects the Norwegian economy. One argument is how the society deals with the dropouts in terms of social security. Another element is that the Norwegian welfare system is in need of this lost share of future human capital. This may be seen as a challenge in the light of the dropout discussion above, which also serve as a source of further discussion.

Competing hypothesis

- *Benchmarking causes too much focus on achieving satisfactory results, which may impact national test score results.*

Although our beliefs support the main hypothesis that benchmarking has positive effects on test score results, we must be aware of possible opposite findings. A competing hypothesis is related to cover opposing outcome, namely that benchmarking may not have positive effect on educational results. This is important in terms of minimizing potential bias, and hence take other considerations into account when analysing our results (Heuer, 2008).

The competing hypothesis may also be supported by the empirical findings on responsiveness to performance information. If the focus on achieving a satisfactory result is too dominant rather than completing the tests as intended, this may cause abuse to the purpose of the tests. Negative information on performance will probably lead to holding the principal and/or municipal head of schooling responsible for the poor results. Clearly, every person in charge favours positive feedback on test outcomes. However, less or no attention may also indicate positive results according to Olsen (2015). Therefore, we find it reasonable to question the surroundings of the tests. Although there is a somewhat

clear understanding that the tests are conducted in order to provide a public overview of average test scores to identify areas of improvements, the consideration of reputation may outweigh this. According to Sørensen and Geys (2017), national tests are designed to customize teaching to their individual needs and improve the standard of education. There is, however, a related discussion regarding elements that may influence these test score results. Therefore, we construct our competing hypothesis to cover such findings.

Literature review

Current literature and research on benchmarking in the public sector are related to a range of public policy reports as well as recent experimental work. We will use both evidence of benchmarking from the private sector, as well as supplementary literature on public policy reforms when investigating our hypothesis. The use of benchmarking in the Norwegian municipal sector is a relatively new implemented practice, which has been introduced to an increasing extent in the municipal sector. As an illustration, the usage of benchmarking in 2008 was present in 26% of the municipalities, while the usage of benchmarking in 2012 increased to 63% (NIBR, 2016). When investigating our hypothesis, we therefore apply evidence on benchmarking from the private sector as well as supplement our beliefs with a range of public policy reports that we find relevant to the topic.

Current literature provides insights on how to implement benchmarking in a business context where the aim is to improve results. 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's paper 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 itself to earlier performance. This may serve as a goal of improvements. When investigating how the use of benchmarking affects national test scores, our estimates is based on how the test scores vary within a municipality, compared to national test scores before benchmarking was implemented.

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 ambitions in a given exercise. The benchmarking literature often refer to aspiration levels when aiming to achieve a common goal. The level of aspiration is often defined at an unreachable level to aim for inspiration. Greve (1998) uses 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 to involve a significant amount of risk. That is, when aiming for a better result, one needs some kind of input that is a necessary to make the desired change. In terms of our hypothesis, this method of benchmarking can serve as a reference when aiming to improve test score results based on own historical results. As we are comparing municipalities with themselves, the decision of implementing benchmarking may involve risk that affects the desired result. Risk may apply when aiming for better test scores through willingness to change, or that public available test scores impact the municipal reputation.

Other related literature involves earlier research on responsiveness to performance in the public sector by Sørensen and Geys (2017), school accountability and performance by Figlio and Loeb (2011), a quasi-experimental estimate on whether class size impact students' performance by Leuven, Oosterbeek and Rønning (2008), and a paper by Propper and Wilson (2003) on performance measurements in public sector among others. When it comes to the methodology part, we elaborate on corresponding methodological literature when presenting our research plan.

Data

Our empirical analysis is conducted using elements from multiple sources of available data. By gathering these data, we create a new dataset where the setup is illustrated in table (1.1), which serve as the basis for further analysis. Our panel data ranges from 2004 to 2016.

We collect national test score data from Skoleporten (Skoleporten, 2017). These are sorted by regions and municipalities based on school averages. National test scores are known as "unadjusted results" that is only available as school average results gathered by each municipality.

Further, we use a school-level test performance as well as municipality-level test performance, both retrieved and conducted by Statistics Norway (SSB, 2017). The school-level test performance is a measure on how each school has contributed to the test score results, compared to general test score results that only measure the individual performance. The indicators are published in a public report by Statistics Norway with the purpose of analysing differences between the particular school and the municipalities' ability to contribute to test score results. The school-level test performance may be interpreted as the result a school would have been given if their student base was average. The municipality-level test performance is the average test score results within a municipality, adjusted for the students' earlier results as well as family background. So far, we only focus on results gathered school by school. However, we might also be given access to test score results on individual level that will be used to support our results.

Additionally, Statistics Norway distinguishes between cross-sectional and value-added indicators. These are captured as a common term on school-level test performance. Cross-sectional indicators consist of information about the students' socioeconomic background, while value-added indicators cover adjustments for the students' earlier test score results. Value-added indicators are only present for test score results in 8^{th} grade, while cross-sectional indicators may be used both in

5th and 8th grade. These indicators may very well correspond to control variables. Therefore, when estimating our models described below, we do not include other control variables when the particular indicators are present in the model. Another important note to make is that the value-added indicator may be seen as each particular schools' contribution to the student's knowledge in between two periods. Therefore, this indicator serves as a good signal on the quality of the school.

The benchmarking data will be prominent as a dummy variable for each municipality; a dummy equal zero prior to implementation of benchmarking and one when benchmarking is implemented. This data is collected through the Ministry of Local Government and Modernization, by Norwegian Social Science Data.

Lastly, we use municipal components form the dataset by Fiva, Hasle and Natvik (2017). Due to the ongoing reform of Norwegian local governments, we use the 2004 municipal structure throughout the analysis (428 units).

Analysis

The data we collect can be set up in the following panel data fashion.

t	***k***	**BM**	**NP5**	**NP8**	**SHE**	**SI**	**P**
2004	101	0	3,6	4,2	0,22	0,08	6700
2004	102	0	4,0	3,8	0,28	0,12	32000
…..	103	1	4,7	4,4	0,39	0,2	55000
2004	2030	1	4,2	4,3	0,19	0,05	2200

Table 1.1: setup

We will have to organize all the available and relevant data in this fashion. This will have to be done for every year where data is available, which is 2004, 2008,

2012 and 2016. For the years in between we summarize the test score results to compute the averages. Then this will be compared to the increase in benchmarking usage.

For each year *t,* we will look at every municipality *k* which has its own number ranging from 101 all the way to 2030. BM is a dummy variable showing if municipality *k* uses benchmarking or not, 0 if not and 1 otherwise. NP5 and NP8 shows the test scores for the 5th graders and 8th graders respectively. SHE is a control variable that looks at the share of students that have parents with a higher degree of education. This is defined as university/college degree or higher. Other control variables are SI, share of students with an immigrant background and P for the population in municipality *k.* The table above is only an exemplification of our potential setup and the numbers in the table are random fictitious examples.

Models

This kind of longitudinal study can contain heterogeneous effects over time. Meaning that other factors affecting national test scores will change over time. The fixed effect method is the key, because it can hold these effects fixed. In turn, we are able to measure the true effect that we are after; namely, the impact of benchmarking on national test scores. When dealing with panel data, where there are longitudinal observations for the same subject, municipalities in this case, fixed effects represent the subject specific means (Angrist & Pischke, 2008). The fixed effect estimator is the estimator for the coefficients in the regression model including those fixed effects.

The variation in our case are time and municipality effects varying over time. This variation may have an impact on the national test scores, which then in turn can complicate our estimation. Because of the fixed effect estimators, these variations are dealt with. When holding the municipality and time variations fixed, their effect on the national test scores are discarded. What are then left with is the estimation of benchmarking on national test scores. BM in our model serves as a dummy variable that is turned on and off for municipality *k* at time *t*, determined by the use of benchmarking or not. To strengthen the models, we also include

several control variables that we believe have a significant correlation with our dependent variable, national test scores. Examples in this model are; share of students with parents who hold a higher education, SHE, share of students with immigrant background, SI, and the population, P. For the population we will take the natural logarithm of it to control for the potential spread in the population affecting the result. In our paper the inclusion or exclusion of such control variables and why they are important will be an important discussion to be made. In this case they are used as examples, variables that potentially could be included. It is important for us that this is more an exemplification of a model setup and not our final version. Hence, it could be that we end up with different, fewer or more control variables. For now, we will just argue that these three intuitively seem like good variables to include and control for. For instance, take the control variable SHE. Parents with a higher academic competence could potentially use that to help their child in achieving higher educational results. Important to mention is that when we use the municipality – and school-level test performance variables, we do not include the control variables as such circumstances are already controlled for in the Statistics Norway-data.

Now that we have argued for the choice of a fixed effect model, and which control variables to include, this is potentially what the fixed effect model could look like:

$$NP8_{kt} = \beta_0 BM_{kt} + \beta_1 SHE_{kt} + \beta_2 SI_{kt} + \beta_3 ln(P)_{kt} + \lambda_k + \theta_t + \varepsilon_{kt}$$

The model above is a version of how the fixed regression could be set up. We will use a fixed effect model because we are dealing with parameters that are characterized as non-random. In contrast to a random effects model in which the group means are a random sample from a population. In these model types, which we will use, each group mean is a group-specific fixed quantity (Ramsey & Schafer, 2002)

The success of this model rests on the assumption that the parameters λ_k and θ_t hold changes in municipality effects fixed over time. Because of this, the model can control for the unobserved heterogeneity as long as this heterogeneity is constant over time. The causal effect of benchmarking on national test scores can

be estimated by treating λ_k, the fixed effect, as a parameter to be estimated. The year effect, θ_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).

Municipality and time effects can affect national test scores, which is something that we do not set out to measure. Examples of such effects are municipality council groupings and other time factors and trends. Another way to put it is that these effects variation variate along with the national test scores. If we have a model with this kind of variation, the effect of benchmarking on national test scores will be difficult to retrieve. When all these factors vary, what is the specific benchmarking variation that affects test scores. Benchmarking variation in this case is simpler, it is either benchmarking or no benchmarking.

Another model that potentially could be included is the following, which looks at the change, Δ, in national test scores for municipality *k*:

$$\Delta NP_k = \gamma BM_k ..... + e_k$$

The effect we are looking for is given by "gamma" when municipality *k* uses benchmarking, BM_k, or not. By differencing the data, the time invariant components of the model are removed. Here, one can measure the effect of the change if municipality *k* implements benchmarking, compared to when they did not (Cameron & Triveldi, 2005).

Organizational differences

An additional factor to look at could be how the different municipalities are organized when it comes to public administration. This can vary from municipality to municipality with regard to responsibility of the educational sector. The organizational setup can differ when it comes to tasks and duties that the local councilman (rådmann) is faced with. It might be that he/she is responsible for schools and education as a whole, or that the municipality in

question has hired a specific head of schooling, who then reports to the councilman. This can affect the level of service towards the local educational sector. To what extent we want to include this in our calibration, models and furthers approach remains to be discussed. However, it is very likely that it will be addressed and analysed in one way or another. There are many interesting theories of organizational structures within the world of the private business sector, as it also is with benchmarking. In the same way it is possible that such theories can be attributed to a public sector perspective.

Conclusion

Our empirical strategy, as described above, serve as the basis for further analysis. We will continue to work by firstly focusing on collecting all relevant data, and then gathering the variables of interests. Although our hypothesis somewhat reveals our belief about the thesis project, we are truly motivated to analyse the models. Hopefully, the outcome will provide valuable contributions to the public sector performance, particularly on how benchmarking as performance measurement serves as a good model of determining educational performance.

References

Angrist, J. D., & Pischke, J.-S. (2008). Mostly Harmless Econometrics. Princeton University Press.

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.

Figlio, David and Susanna Loeb. 2011. School Accountability. In Handbook of Economics of Education (Volume 3), edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland.

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

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

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

Huer, Richards J. Jr, (2008): "Chapter 8: Analysis of Competing Hypotheses", Psychology of Intelligence Analysis, Center for the Study of Intelligence, Central Intelligence Agency

Kunnskapsdepartmentet (2008-2009): 2 En variert og mer praktisk grunnopplæring. Utdanningslinja. (St.meld. nr. 44 2008-2009). Retrieved from https://www.regjeringen.no/no/dokumenter/stmeld-nr-44-2008-2009-/id565231/sec2

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

Nielsen, Poul and Donald P. Moynihan. 2017. How Do Politicians Attribute Bureaucratic Responsibility for Performance? Negativity Bias and Interest Group Advocacy. *Journal of Public Administration Research and Theory*, 27(2): 269-283.

Nordlys. (6.8.2010): "Frafall er et forutsigbart skoleresultat". Retrieved from https://www.nordlys.no/kronikk/frafall-er-et-forutsigbart-skoleresultat/s/1-79-5217139

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

Olsen, A. L. (2015). Negative Performance Information Causes Asymmetrical Evaluations and Elicits Strong Responsibility Attributions.

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

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, SSB (2017): "Er det forskjeller I skolers og kommuners bidrag til elevenes læring I grunnskolen? – En kvantitativ studie" retrieved from https://www.ssb.no/utdanning/artikler-og-publikasjoner/hvor-mye-bidrar-skoler-til-elevers-laering

SINTEF teknologi og samfunn (2007). "Tiltak mot frafall i videregående opplæring" Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Tiltak-mot-frafall-i-videregaende-opplaring-2007/

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