Pursiainen, Heikki; Kortelainen, Mika; Pääkkönen, Jenni

Conference Paper

Impact of School Quality on Educational Attainment - Evidence from Finnish High Schools

54th Congress of the European Regional Science Association: "Regional development & globalisation: Best practices", 26-29 August 2014, St. Petersburg, Russia

Provided in Cooperation with:
European Regional Science Association (ERSA)

Suggested Citation: Pursiainen, Heikki; Kortelainen, Mika; Pääkkönen, Jenni (2014) : Impact of School Quality on Educational Attainment - Evidence from Finnish High Schools, 54th Congress of the European Regional Science Association: "Regional development & globalisation: Best practices", 26-29 August 2014, St. Petersburg, Russia, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
https://hdl.handle.net/10419/124369

Impact of School Quality on Educational Attainment - Evidence from Finnish High Schools

Mika Kortelainen*, Jenni Pääkkönen, Heikki Pursiainen†

July 24, 2014

We analyze differences in school quality using a comprehensive data set covering all upper secondary school graduates in Finland during the years 2002-2013. School quality is defined as the effect of the school on matriculation exam results controlling for quality of student intake. In other words, the quality difference between two schools is the expected difference in exam results for a randomly chosen student switching schools. Using methods similar to Chetty, Friedman and Rockoff (2013) we are able to measure both cross-sectional differences in school quality and the persistence of these differences over time. We also control for the uncertainty inherent in assessing the quality of smaller schools with a relatively low number of graduates. We use each pupil's comprehensive school grades to control for previous education / pupil quality. Also, comprehensive school fixed effects are used to control for differences in comprehensive school grading as well as unobserved socioeconomic factors. The method is potentially sensitive to bias induced by school selection. To assess the potential bias we partially match our student sample to a spatial database by home address and use this to assess bias. We find no evidence of significant bias. Our first result is that there are significant cross-sectional differences in school quality even after controlling for student intake quality. The quality difference between the top schools and bottom schools each year measured in average matriculation score points is around one grade point in a scale of 1 to 7. In Finland university entry is partly controlled by these matriculation exam results. A one-point difference in grade averages will significantly affect the chances of entry into the most competitive university curricula. This result must, however, be qualified in a number of ways. First, large differences are observed only between the very

*Finnish Government Institute for Economic Research

†Finnish Government Institute for Economic Research, Arkadiankatu 7, 00101 Helsinki. heikki.pursiainen@vatt.fi

top and bottom institutions. Most schools are much closer to each other in quality: the interquartile range each year is only about a fifth of a grade average point. Most schools are thus clustered quite close to each other in quality. Also, while there is persistence over time in school quality, this is far from complete. This means that the ranking of the middling-quality majority of schools is highly unstable over time, making any yearly league tables highly suspect. There is more persistence in the very top and bottom institutions. Finally, school quality seems to be for the most part evenly distributed regionally. While there are certainly good schools in the largest cities, the success of the most selective institutions is mostly explained by quality of intake rather than teaching.

1 Introduction

The existence and size of quality differences between upper secondary schools is a topic of some debate in Finland. Upper secondary schools choose pupils based on their grades in lower secondary school, so that at least in the largest metropolitan areas best pupils tend to cluster into "elite" upper secondary schools. Pupils graduate from upper secondary schools by taking a nationwide matriculation exam. Entrance into universities is partly controlled by the results of these matriculation exams. For the most competitive university curricula small differences in matriculation can significantly affect chances of entry.

This system, which is described in more detail in Section 2, means that quality differences between upper secondary schools matter. The appearance of new matriculation results is a matter of considerable interest in the media and among the public in general. Numerous rankings or league tables are published purporting to identify the best upper secondary schools (USS) in the country. However, these league tables are not based on solid statistical methods. Some of them make no at all effort to control for the quality of student intake. Still others use dubious *ad hoc* methods to try and control for student quality.

In this paper we use statistical methods to measure quality differences between upper secondary schools and explore whether these can be translated into meaningful league tables or rankings. Quality differences between schools are defined to be differences in matriculation exam success after controlling for the quality of student intake. Such measures are usually called value-added measures in the literature. There exists a large literature on how to use value-added methods to reliably compare schools or teachers, (see for example the surveys [10, 21]).

Based on this literature, we identify some properties that a reasonable value-added model of USS quality should possess. First, it must control for the quality of pupil intake and selection as well as possible. Another important thing is to account properly for school quality change in time. Obviously there will be persistence in USS quality, as for example teacher composition varies only slowly. But as our period of interest is over a decade long, allowance must be made for the possibility of significant quality

change over time. A good value-added model also takes into account the uncertainty concerning school quality estimates. This is especially important for small schools, of which there are relatively many in Finland. Yearly quality estimates for small schools are very uncertain because of small sample size. Usually this uncertainty is taken care by using estimators with the so-called shrinkage property. Shrinkage means that the more uncertain a school-quality estimate is, the more the estimate gets shrunk towards the mean.

We estimate a value-added model which satisfies all these requirements. Out of the many possible models we choose one recently proposed by Chetty, Friedman and Rockoff [4, 5]. A value-added measure is calculated for every (daytime) USS in Finland for the period 2002-2013.

The results are mixed. Our first result is that there are significant cross-sectional differences in school quality even after controlling for student intake quality. The quality difference between the top schools and bottom schools each year measured in average matriculation score points is around one grade point in a scale of 1 to 7. In Finland university entry is partly controlled by these matriculation exam results. A one-point difference in grade averages will significantly affect the chances of entry into the most competitive university curricula.

This result must, however, be qualified in a number of ways. First, large differences are observed only between the very top and bottom institutions. Most schools are much closer to each other in quality: the interquartile range each year is only about a fifth of a grade average point. Most schools are thus clustered quite close to each other in quality.

Also, while there is persistence over time in school quality, this is far from complete. This means that the ranking of the middling-quality majority of schools is highly unstable over time, making any yearly league tables highly suspect. There is more persistence in the very top and bottom institutions, which are roughly the same during the whole period under consideration.

Finally, school quality seems to be for the most part evenly distributed regionally. While there are certainly good schools in the largest cities, the success of the most selective institutions is mostly explained by quality of intake rather than teaching.

The paper is structured as follows: in the next section we decribe the Finnish educational system. Section 3 gives a brief review of some relevant literature, Section 4 describes the method used, while 5 decribes the data. The main results are presented in Section 6 and their reliability and bias are discussed in Section 7. The last section , Section 8 includes a brief final discussion of the relevance of the results.

2 The Finnish school system

In Finland education is compulsory for everybody between seven and sixteen years of age. In practice, almost everybody enters comprehensive school at age 7 and graduates at age sixteen. Comprehensive school consists of primary school (grades 1 to 6) and lower secondary school (grades 7 to 9). After graduation from comprehensive school around 50 % of each cohort enters an upper secondary school (Below, if we want to be particularly

precise, we use the acronym USS for upper secondary school. When no confusion can arise, an USS is referred to simply as a school.) Most of the rest enter vocational training.

Entry into upper secondary schools is competitive and based on comprehensive school success. In smaller towns, there is often only one or two upper secondary schools, and in practice everybody who wants to enter one, gets in. However, in larger cities, there are "elite" upper secondary schools entry into which requires top comprehensive school grades.

The basic curriculum in an USS consists of general knowledge subjects such as Finnish, foreign languages, mathematics and various human and natural sciences. Studies in each of these is broken up into courses. Graduation from an USS requires succesful completion of at least 75 courses. Pupils have substantial freedom of choice as regards which courses to take. However, there are some compulsory ones. These include Finnish, Swedish, a foreign language (in most cases English) and basic mathematics.

Graduation from an USS requires passing a matriculation examination in addition to completing the required courses. Exams have to be passed (roughly speaking) in at least four different subjects. Of these the Finnish (or Swedish for the Swedish-speaking minority) exam is compulsory. The other three can be chosen from a choice set determined by the courses completed during USS studies. The matriculation examination takes place in each spring and autumn. Traditionally, candidates would take their exams in the end of the spring term of their 3rd year, but the modern system is somewhat more flexible. Typically, students now take the exams in two or more parts. Most still graduate after three years, but there are people who graduate after only 2 years, as well as some who take 4 years to graduate.

Success in upper secondary school is a major determinant of university entry. An overwhelming majority of university entrants are USS graduates. Entry into university is based partly on USS and matriculation exam grades, partly on entrance exams. Some university departments, such as mathematics and physics do not require participation in entrance exams for candidates with top grades in relevant matriculation exams. For those university curricula that do require entrance exams, the entry threshold in these exams in significantly lower for candidates with good matriculation exam results. This means that possible differences in USS quality may translate into significant differences in entry into the most competitive university curricula, such as medicine and law.

3 Related literature (to be expanded)

There exist a large literature on teacher and school evaluation. Of these two, teacher evaluation has perhaps been the focus of more interest at least among economists. One reason for this is the rather heated debate concerning teacher tenure and remuneration going on in the US. Most of the evaluation literature concerns different value-added methods based on success in standardized exams. Value added is loosely defined as the the contribution of the teacher or school on top of each students initial level of accomplishment. Value-added methods therefore use statistical techniques to control for the initial "quality" of student intake. For surveys of the value-added literature, the

reader is referred to for example [10, 16, 21, 14].

In addition to controlling for initial student quality, most value-added methods try to take into account the uncertainty inherent in school / teacher evaluation. In most cases, each class or cohort of students in a school is rather small. This means that year-to-year variation in outcomes may be quite large. A standard way of accounting for such uncertainty is to use so-called shrinkage estimators [23, 13, 4].

A shrinkage estimator shrinks uncertain value-added estimates towards the mean. The more uncertain the estimate, the closer to the mean it is shrunk. Shrinkage is quite important when one compares the performance of Finnish USS, as many of these are small, with only one class of perhaps 20-30 students or even less graduating each year. If school performance is evaluated without taking the uncertainty resulting from small cohorts into account, the "best" and "worst" schools each year tend to be the smallest ones. This is evident from the unscientific yearly USS rankings published by Finnish newspapers [1, 28].

A significant part of the recent value-added literature has been about the time-series properties of value-added estimates. More precisely, this literature has focused on the persistence in school / teacher quality in time. Traditionally, value-added methods have assumed that school quality remains constant over time. This assumption has been tested in the more recent literature, and the results show that while value-added estimates show significant persistence, this is far from perfect and leaves room for time series variation, see for example [2, 4, 11, 7, 22].

Perhaps the most difficult problem facing value-added estimation is the one posed by selection. In our case, selection is a problem if students self-select into upper secondary schools based on (to us) unobserved properties, such as motivation. If such self-selection is significant, it can result in biased value-added estimates, as these capture the effect of the unobserved variation as well as the effect of the school. For example, Rothstein [26, 27] considers selection to be a significant, perhaps even invalidating problem for value added estimation. This assessment has however, come under rather severe criticism, see for example [6, 17, 19]. Attempts have been made to measure bias caused by selection by using randomized experiments[15, 12]. The results give some evidence against significant selection bias.

In the main methodological source for the current paper [4], the authors attempt to quantify selection bias using a variety of methods, some of which we also implement. The authors do not observe significant selection bias.

Also, there is some evidence that value-added estimates predict future labour market success and / or selection into tertiary education [5, 3]. This would seem to indicate that value-added measures actually describe something that has real significance.

In addition to the general value-added literature, this paper is also based on previous literature concerning the Finnish upper secondary school system. This is by no means a large literature, perhaps the most prominent one is [18]. Others include [20, 25].

4 Method

4.1 General remarks on value-added methods

As mentioned, the methods used in this paper belong to the class of value-added methods. This means that we attempt to take into account the fact that different upper secondary schools often have very different students. Because students are different to begin with, it comes as no surprise that matriculation exam results differ widely between schools. A good result from a student who was already succesful in comprehensive schools is not at least completely due to her upper secondary school. Similarly a mediocre result from a student who was a bad student in comprehensive school may be counted as a succes for the upper secondary school. A value-added model tries to formalize this intuition by attempting to control statistically the initial quality of each school's student intake. There are many different ways of doing this, but all share the goal of isolating the contribution of the school to the examination results from the contribution of student intake. Different models produce somewhat differing results, and as the literature review of the previous section demonstrates, the debate over the relative merits of different models has been active.

In the next subsection we present the methods used in this paper. It is a rather sophisticated value-added method that addresses for example the problem of persistence of school quality. It is however important to keep in mind from the beginning that even the most sophisticated of value-added methods are vulnerable to the selection problem. This problem is discussed in more detail in subsection 4.3.

4.2 The value-added model

In this subsection we describe briefly the method used. It is based to a large extent on articles [4] ja [5], in which it is described in more detail. Each student i, matriculationg in the year t is in some upper secondary school $j = j(i,t)$. Her / his score in the matriculation exam is $A^*_{i,t}$, which is determined according to the following equation:

$$A^*_{i,t} = \mu_{j,t} + \beta' X_{i,t} + \varepsilon_{i,t}. \tag{1}$$

Here $\mu_{j,t}$ is the effect (true value-added) of school j in year t, $X_{i,t}$ is a collection of covariates describing the characteristics of student i matriculating at time t. There are K of these, β is the coefficient vector associated with these covariates and ja $\varepsilon_{i,t}$ is the idiosyncratic error term of student i.

The most interesting part of equation (1) is naturally the school effect $\mu_{j,t}$. It describes the improvement or deterioriation caused by the upper secondary school on the student's exam results. This effect may be understood using a thought experiment. Suppose we were to choose a number of students by random sampling and to assign these students randomly to schools A and B. Then the difference of their matriculation exam results would be $\Delta_{A,B,t} = \mu_{A,t} - \mu_{B,t}$.

For example if the true effect of school A in the year 2012 is $\mu_{A,2012} = 0.4$ and the effect of school B is $\mu_{B,2012} = -0.5$, this means that the expected exam results of the

randomly assigned students would be 0.9 grade points higher in school A in year 2012. The effect of the average school is normalized to 0.

For the method used to be valid, the terms of equation (1) must satisfy certain assumptions, of which the first are:

1. The school effect process $\mu_{j,t}$ for each school j follows the same zero-mean covariance stationary process and is also otherwise well-behaved enough.

2. The joint distribution of the zero-mean error terms $\varepsilon_{i,t}$ and the school effects is time-homogeneous[1]. Error terms and school effects are not correlated between schools.

These assumptions have an intuitive interpretation. They are about the variation of the school effect in time. The model allows for the school effect to change in time. This might mean for example, that the effects of top and bottom schools revert towards the mean as time goes on. How rapidly such change occurs is an empirical matter and is estimated from the data. The effects of schools might decay rapidly, or might be very persistent over time. What the model does not allow is for the school effects to increase or decrease without limit, or for the rate of change to vary over time. These are rather mild requirements considering the sample and time period we are looking at.

These assumptions guarantee that the value-added measure we use is a well-defined best linear prediction (to be constructed below). They emphatically do not guarantee that our value-added measure is an estimator of the "true" school effect. In particular, these assumptions allow for the school effects to be correlated with the error terms as well as the error terms with each other. Each of these correlations may result in bias in the value-added measures as estimates of the school effect. A prominent reason for why such correlations could exist is self-selection. We will discuss this below.

As mentioned, these assumptions allow us to define the value-added measure we use. It will be calculated in three steps.

First, the results of the matriculation exam $A^*_{i,t}$ are regressed on the covariates $X_{i,t}$ and school fixed effects[2]. This gives an estimate $\hat{\beta}$ for coefficients β. Using the estimated coefficient we may calculate the residualized exam score $A_{i,t} = A^*_{i,t} - \hat{\beta}' X_{i,t}$. If $\hat{\beta}$ is well estimated, the equation $A_{i,t} = \mu_{j,t} + \varepsilon_{i,t}$ holds approximately. The residualized score is then the exam score purified of the effects of the covariates, or put otherwise, the exam score after controlling for the initial quality of the student. When 1 ja 2 hold, the distribution of the residualized score is quite simple.

In the next stage, the mean residualized score is calculated for each school-year combination. The resulting variable is $\overline{A}_{j,t} = \frac{1}{n_{j,t}} \sum_{i=1}^{N_{j,t}} A_{i,t}$. Here $n_{j,t}$ is the number of

[1] More precisely, the variances of the error terms are constant for each year and all schools. If the error terms are correlated with each other and / or the school effects, these correlations are constant for all schools and years. Correlations between terms in different years depend only on the length of the period separating them.

[2] The method gives a good estimate of β even when the covariates $X_{i,t}$ are correlated with the school effects. In principle it is possible, that the changes in school effects are correlated with the covariates, in which case the estimation procedure would fail. This is, however, probably not a big problem, as even dropping the fixed effects entirely did not significantly change the residualised scores.

matriculating students in school j in year t and the sum is over all i with $j(i,t) = j$. The equation $\overline{A}_{j,t} = \mu_{j,t} + \overline{\varepsilon}_{j,t}$ holds approximately with $\overline{\varepsilon}_{j,t}$ being the mean error term in school j in year t.

This allows us to construct the value-added measure. The estimator recommended by [4] is the best linear prediction of the mean residualized score of a school conditional on all previous / other mean residualized scores, or $\hat{\mu}_{j,t} = \hat{E}(\overline{A}_{j,t}|\overline{A}_{j,t-1}, ..., \overline{A}_{j,t-s})$. This formula and all calculations below are presented in the interest of simplicity for the situation in which we are predicting the result of school j in year t with the s previous years[3]. Conditional on assumptions 1 and 2 this linear prediction is of the form:

$$\hat{\mu}_{j,t} = \hat{\Sigma}_{j,t}^{-1}\hat{\gamma}_{j,t},$$

in which $\hat{\Sigma}_{j,t}$ is the estimated $s \times s$ covariance matrix of $\overline{A}_{j,t-1}, ..., \overline{A}_{j,t-s}$ and $\hat{\gamma}_{j,t}$ is the $s \times 1$ covariance vector of $\overline{A}_{j,t-1}, ..., \overline{A}_{j,t-s}$ with $\overline{A}_{j,t}$.

When the assumptions are met, the required covariance estimates may obtained as follows. The covariance $Cov(\overline{A}_{j,t}, \overline{A}_{j,t-k}) = \sigma_{\overline{A},k}$ depends only in the lag k and is constant for all schools j. The autocovariances $\sigma_{\overline{A},k}$ may be estimated as a mean of the intra-school autocovariances of the mean residualized scores, weighted by the number of students. These estimates form the off-diagonal elements of the matriz $\hat{\Sigma}_{j,t}$, by setting $\left[\hat{\Sigma}_{j,t}\right]_{k,l} = \hat{\sigma}_{\overline{A},|k-l|}$ when $k \neq l$. Also the vector $\hat{\gamma}_{j,t}$ is formed using these estimates, so that $[\hat{\gamma}_{j,t}]_k = \hat{\sigma}_{\overline{A},k}$.

In addition to these we need an estimate for the variances $Var(\overline{A}_{j,t-k})$. A suitable estimator is given by

$$\hat{\sigma}^2_{\overline{A}_{j,t}} = \hat{\sigma}^2_A - \hat{\sigma}^2_\varepsilon + \frac{\hat{\sigma}^2_\varepsilon}{n_{j,t-k}},$$

where $\hat{\sigma}^2_A = \frac{1}{N-K-1}\sum\sum(A_{i,t} - \overline{A})^2$ and $\hat{\sigma}^2_\varepsilon = \frac{1}{N-K-C-1}\sum_j\sum_t\sum_{j(i,t)=j}(A_{i,t} - \overline{A}_{j,t})^2$. In these K is the number of covariates and C the number of school-year combinations while $n_{j,t-k}$ is the number of matriculating students in school j in year $t-k$. These estimates form the diagonal elements of $\hat{\Sigma}_{j,t}$ so that $\left[\hat{\Sigma}_{j,t}\right]_{k,k} = \hat{\sigma}^2_{\overline{A}_{j,t-k}}$. Now we have estimated all the elements required to calculate $\hat{\mu}_{j,t}$.

4.3 The selection problem

When assumptions 1-2 hold the resulting estimator is the best linear prediction $\hat{\mu}_{j,t}$ for the mean residual score $\overline{A}_{j,t}$ in school j in year t:lle conditional on the residual scores of previous years. But we are more interested in estimating the actual school effect instead of the mean residual score. When is the value-added estimator $\hat{\mu}_{j,t}$ an estimator of the true school effect? Additional assumptions are needed for this. For example, the following two assumptions are sufficient:

3. The school effect $\mu_{j,t}$ is not correlated with the idiosyncratic shocks $\varepsilon_{i,s}$.

[3] In our real calculations there is not data for all school-year combinations. We also calculate predictions for each year in the data conditional on all other years in addition to the prediction for the last one given all previous years.

4. The idiosyncratic effects $\varepsilon_{i,s}$ are uncorrelated with each other.

If these assumptions hold in addition to the previous ones, we have

$$\hat{\mu}_{j,t} = \hat{E}(\overline{A}_{j,t}|\overline{A}_{j,t-1}, ..., \overline{A}_{j,t-s}) = \hat{E}(\mu_{j,t}|\overline{A}_{j,t-1}, ..., \overline{A}_{j,t-s}), \tag{2}$$

so that the value-added estimator is also the best linear estimator for the school effect given the information of previous years.

These assumptions also have an intuitive interpretation. The assumptions rule out selection into schools based on unobserved student characteristics. Students are of course selected into schools based on their characteristics, chief among these their grades in comprehensive school. This is not a problem, as we can control for such selection by including the comprehensive school grade point average in the vector of characteristics $X_{i,t}$. The selection problem arises if students select into institutions based on unobservable characteristics such as study motivation.

Consider students from a certain comprehensive school choosing either upper secondary school A or B. Suppose that there are two groups of students, with on average identical comprehensive school grades, but which differ on the level of motivation. Of these two schools, school A offers a more rigorous course of study, thus attracting the more motivated students in the first group. School B is more laid back and attracts the less motivated students in the otherwise identical second group. Because of the more rigorous approach, school A produces better results than school B. But students in school A benefit also from their superior motivation. Thus exam results in school A are better not just because the school is better, but also because of more motivated students. In this case, the value-added estimator exaggerates the effect of school A compared to school B, because from the point of view of the researcher, the two groups of students appear identical.

Technically, such selection can be described as correlation between the student shocks $\varepsilon_{i,t}$ which include motivation and the school effects $\mu_{j,t}$. In this example, the correlation is positive. Using more convoluted examples it is perhaps possible also to produce negative correlations, resulting in underestimation of school effects. Note that the problem persists even if schools A and B have identical "true" effects. If one of these consistently attracts more motivated students, its estimated value added will appear larger. Thus, even certain types of correlation between the idiosyncratic effects $\varepsilon_{i,t}$ will also produce biased estimates, thus we need some assumption like Assumption 4.

It seems that Assumption 4 could be weakened somewhat, but it is unclear of how much use this is. Consider for example peer effects. Suppose that the exam results of students are affected by their peers. If your classmates are exceptionally motivated, this will help you as well. This may be modeled as correlation between the idiosyncratic effects. Suppose further that there is no selection problem, students are selected randomly (with regard to their unobserved characteristics). It's just that some years, the students are more motivated on average and some years less motivated on average. Assume also that a student's peer group consists only of the students matriculating at the same time, so that the idiosyncratic effects are correlated only with the idiosyncratic effects of the same year, not with either school effects or idiosyncratic effects in other years. This highly

specific form of correlation seems not to cause systematic bias in the estimator described above. It is, as mentioned, a highly specific form of correlation that seems unlikely to be an accurate description of any realistic peer effects. Also, while this form of correlation does not invalidate the estimator described above, it will invalidate the other estimator used in this paper which is described in the next subsection.

To sum up this discussion, selection poses a potentially fatal problem for value-added estimation. If students self-select into schools based on their unobservable characteristics, value-added estimators are biased with respect to true school effects. This is basically true of any value-added estimator, not just the ones used here. While it is perhaps impossible to provide estimates fully free from doubt of selection bias, there exist some approaches that can be used at least to attempt to detect it. In this paper we adopt an approach similar to [4]. In other words, we proceed to calculate our value-added estimates as if Assumptions 3-4 were satisfied. We then submit these estimates to various tests that attempt to measure the amount of bias. These tests are discussed below in Subsection 4.6.

4.4 Some alternative value-added measures

The estimator described above is based on [4] and [5]. This estimator uses information from all other years except t to predict the effect on year t. This type of jackknife estimator which does not use information of the "current" year, is suited to the purposes of these articles. The estimator is used for various kinds of predictive purposes. When using value-added estimators for prediction, it is necessary to discard information of the period for which one is predicting. Otherwise the same estimation error is introduced on "both sides" of the predictive equation and misleading correlation is thus guaranteed.

In this paper whenever we do predictions similar to [4], we use the jackknife estimator. However, in contrast to that paper, we are also interested just in describing the quality differences between Finnish upper secondary schools. As these estimates are not used in any further analyses, we may use estimators that use all available information, including scores from the year whose value-added we are calculating. When assumptions 1-4 hold, it is possible to calculate the best linear prediction $\tilde{\mu}_{j,t} = \hat{E}(\mu_{j,t}|\overline{A}_{j,t}, ..., \overline{A}_{j,t-s})$, in other words, predict the school effect conditional on all available exam scores, including from the year of interest t. This prediction can be construed in a fashion similar to the jackknife estimate by changing the covariances in the estimator equation slightly.

The estimators described so far allow the school effect to change over time. If the school effects can be assumed to be constant, the estimators become simpler. Without going into details, the covariance matrices and vectors described above simplify and result in the following estimators, which we call Kane-Staiger estimators after [15]. The jackknife estimator becomes

$$\hat{\mu}_{j,t}^{KS} = \overline{A}_j^{-t} \frac{\hat{\sigma}_A^2 - \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_A^2 - \hat{\sigma}_\varepsilon^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_{j,t-1}+....+n_{j,t-s}}} \tag{3}$$

and the estimator based on all the data becomes

$$\tilde{\mu}_{j,t}^{KS} = \overline{A}_j \frac{\hat{\sigma}_A^2 - \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_A^2 - \hat{\sigma}_\varepsilon^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_{j,t}+....+n_{j,t-s}}}. \tag{4}$$

Here $\overline{A}_j^{-t}$ is used to describe the average of all residualized exam scores in school j except the ones from year $t$and $\overline{A}_j$ is the mean when these results are also included.

4.5 The shrinkage property

Equations (3) and (4) show clearly that these are so-called shrinkage estimators. The mean of the residualized scores is shrunk towards zero. The amount of shrinkage is determined by the uncertainty concerning these means. If there is a lot of uncertainty measured by the estimated individual-level variance and the inverse of the number of observations used to calculate the mean, the amount of shrinkage is also large. All things being equal, the estimates for small schools are shrunk more than estimates for larger schools. Without this shrinkage, estimates for small schools will be highly variable.

The shrinkage property can be seen directly for the Kane-Staiger type estimators. While it is less obvious from the structure of our more complicated estimators $\hat{\mu}_{j,t}$ and $\tilde{\mu}_{j,t}$, they also have this shrinkage property. Thus, these are shrinkage estimators with the additional property of allowing school quality to change over time.

4.6 Forecast bias

As already mentioned, selection can induce systematic bias in estimation in the sense that the value-added estimators are not concentrated near the "true" school effects. Technically speaking, the existence of bias results from violations of Assumptions 3 and 4. It has already become clear that there are many possible reasons for such violations. It is not an easy task to find out whether there is bias. It is even more difficult to quantify the amount of bias. We do not even attempt a comprehensive discussion of bias. We concentrate on a particular type of bias, called forecast bias in the literature [4].

The definition of forecast bias has to do with the already familiar randomization thought experiment. The true school effects measure the differences in exam scores in a randomized experiment in which students are randomly assigned into schools. Value-added estimators are forecast unbiased if the value-added measures on average correctly predict school differences in such a randomized experiment.

Suppose that jackknife value-added estimates $\hat{\mu}_{j,t}$ are available for all schools. Now randomly assign students to schools and estimate the regression model

$$A_{i',t} = \alpha + \lambda \hat{\mu}_{j,t} + \nu_{i',t}$$

for all randomized students i'. The notation i' is used to emphasize, that student i' is part of the randomized experiment and did not participate in the usual selection process. The value-added estimator is forecast unbiased if $\lambda = 1$. The amount of forecast bias is given by $B(\hat{\mu}) = 1 - \lambda$. Using the definition of the regression coefficient we get

$$\lambda = \frac{Cov(A_{i',t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} = \frac{Cov(\mu_{j,t} + \varepsilon_{i',t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} = \frac{Cov(\mu_{j,t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})}$$

, where the last equation results from randomization.

Forecast unbiasedness implies that estimated value-addeds predict school effects correctly on average. If a policy increases value-added by 0.1 grades, it can be predicted that the resulting increase in grades will also be 0.1. Forecast unbiasedness does not, however, guarantee that each and everyone of the school value addeds is correctly estimated. It is possible that a forecast unbiased method produces wrong estimates for individual schools. Forecast unbiased estimators have some policy relevance, as changes in value added result on average in similar changes in performance. It is, however, to be emphasized that even if forecast unbiasedness can be plausibly demonstrated resulting school rankings or league tables should be regarded with caution.

As the example of motivation clearly shows, it is not easy to quantify selection on unobservables. It is difficult to find direct measures of study motivation. One possible solution is to control for more student characteristics. The characteristics that first suggest themselves are various socioeconomic variables of students and their parents. In a nutshell the idea is that the more observables are included in the residualization model, the less room for unobservables there is. For example, if motivation is correlated with socioeconomic characteristics, the selection problem caused by motivation is alleviated if these characteristics are controlled for. The approach has some obvious weaknesses, but it is mostly the best we can do.

As is explained in the section describing our data, we have access to only a limited number of student-level variables. A subsample of students had information on their address at the time they applied for upper secondary school. These addresses could be linked to the Grid Database of Statistics Finland. The database contains information on averages of socioeconomic variables in the locality that the students come from. The data is described in more detail in Section 5. Whenever we refer to socioeconomic variables we are talking about these local averages.

How can this socioeconomic data be used to assess forecast bias? As already mentioned, the idea is that if for example motivation is correlated with socioeconomic variables, then controlling for these will reduce selection bias caused by motivation. Testing for forecast bias is based on this same idea. If there is significant bias, including new control variables should reduce it insofar the new variables are correlated with the source of bias. Therefore if the value-added estimates are significantly changed by including new control variables, this causes concern. In contrast, if the value-added estimates remain practically unchanged, no evidence of bias is found.

Let's look at this more precisely. If Assumptions 1 and 2 are satisfied, the value-added is the best linear predictor for residualized exam scores in school j . Therefore for student i matriculating from school j in year t has

$$1 = \frac{Cov(A_{i,t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} = \frac{Cov(\mu_{j,t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} + \frac{Cov(\varepsilon_{i,t}, \hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})},$$

which means that forecast bias is equal to $B(\hat{\mu}) = \frac{Cov(\varepsilon_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})}$.

Suppose now, that in addition to the covariates $X_{i,t}$ used in value-added estimation new covariates $P^*_{i,t} = (P^1_{i,t}, ..., P^K_{i,t})$ become available. These could be for example socioeconomic variables as discussed above. Assume further, that these covariates should be included in the "true" value-added model, so that $A^*_{i,t} = \mu_{j,t} + \beta' X_{i,t} + \rho' P^*_{i,t} + u_{i,t}$. Because the new variables were ignored in the original estimation procedure, the resulting residualized scores are of the form (ignoring estimation error) $A_{i,t} = \mu_{j,t} + \rho'(P^*_{i,t} - \hat{P}_{i,t}) + u_{i,t}$. In this equation each component of $\hat{P}_{i,t}$ is of the form $\hat{P}^k_{i,t} = \nu^{k\prime} X_{i,t}$, with ν^k being coefficients from the regression of $P^{*k}_{i,t}$ on school fixed effects and original covariates X.

Define now $P_{i,t} = P^*_{i,t} - \hat{P}_{i,t}$. The original individual effect $\varepsilon_{i,t}$ can now be decomposed into two components: $\varepsilon_{i,t} = \rho' P_{i,t} + u_{i,t}$. First of these is the part correlated with the newly available covariates, while the second component $u_{i,t}$ is uncorrelated with these.

Assume now, optimistically, that selection happens only based on the new variables $P_{i,t}$. In this case the amount of forecast bias is $B(\hat{\mu}) = \frac{Cov(\varepsilon_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} = \frac{Cov(\rho' P_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})}$. This can be estimated in a quite straightforward fashion. First, construct variables $P_{i,t}$. Estimate a model for each $P^{k*}_{i,t}$ regressing it on school fixed effects and covariates $X_{i,t}$. Use these models to form residualized variables $P^k_{i,t}$ using a similar procedure that was used to construct $A_{i,t}$. Then estimate a model regressing $A_{i,t}$ on these new variables and school fixed effects. In this way an estimate $\hat{\rho}$ is formed for the coefficients ρ and thus it is possible to construct the prediction $A^p_{i,t} = \hat{\rho}' P_{i,t}$. Forecast bias is now estimated with a regression model that regresses $A^p_{i,t}$ on the value-added estimates $\hat{\mu}_{j,t}$.

It is possible to find a similar estimate in a more direct route. First, calculate new residualized exam scores with a regression model containing both $X_{i,t}$ and $P^*_{i,t}$ in addition to school fixed effects. Call these new residualized exam results $A^N_{i,t}$. For the "old" and new residualized scores we have approximately $A_{i,t} - A^N_{i,t} = \rho' P_{i,t}$. But this implies that forecast bias is $B(\hat{\mu}) = \frac{Cov(\rho' P_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} = \frac{Cov(A_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})} - \frac{Cov(A^N_{i,t},\hat{\mu}_{j,t})}{Var(\hat{\mu}_{j,t})}$. In other words, forecast bias is estimated by $\hat{B}(\hat{\mu}) = \hat{\nu} - \hat{\nu}^N$, where $\hat{\nu}$ is the regression coefficient from a model predicting the original residualized scores using value addeds and $\hat{\nu}^N$ is the corresponding coefficient from a model, in which the explained variable is the new residualized score.

It is possible to assess forecast bias using additional control variables. It is, however, important to notice that this approach will capture all of the bias only if selection is based only on the additional variables. If this assumption is violated, there can be forecast bias even though this method produces no evidence of bias. In other words, the method necessary produces no evidence of bias when there is none, but bias could exist even when no evidence is provided by it.

5 Data

The data used is mostly combined from two sources. The first source is the database of the Finnish Matricularion Examination Board, consisting of all matriculation exam results from 1990 to 2012. The second source is the National Board of Education database,

which has data on (nearly) all upper secondary school applications from 1998 to 2012. The records in the two databases were combined using the unique national identification number of each student. The time period of study was selected based on the temporal overlap of the two data sources.

The matriculation exam data contains information of all exams taken by each student / candidate, the time of each exam taken and on the grade or failure to pass. There are around 5 million exam results in different subjects in the original database. In addition to information already mentioned, the database contains variables describing the type of candidate, ie. whether the student is a normal candidate, a re-taker etc. The database does not contain information on the date of matriculation or whether the exam was compulsory for the student taking it. There have also been some changes in regulations during the period of interest. For these reasons it was impossible to say with complete certainty when each student graduated. Because of this, a matriculation date for each student was constructed from the data based on simple rules.

The rules used were as follows. First, all students that had at some point abandoned the matriculation process and then restarted were removed from the database. Such students were rare and it would have been difficult to assign these students to a single school. Also, all students that had exam codes that could not be resolved assigned to them were removed. There are examinations both in spring and autumn. The autumn examinations are different from the spring ones, and most students choose to graduate in spring. For these reasons all analyses were carried out on an annual instead of a semiannual level. Each student was assigned one spring as her / his graduation date. This date was chosen to be the first spring that the candidate had passed exams in at least four subjects. Passing was defined as either getting a passing grade for all four or getting an overall passing grade taking into account the compensation rules pertaining to the matriculation exams.

The grade average for each passing student was calculated to be the average grade of the best four exams taken up until and including the graduation date. The grades in the Finnish matriculation are actually Latin phrases, but the Matriculation board converts these into numbers using a scale in which the best grade is translated into 7, the second best into 6, etc. The lowest passing grade is converted to 2 and fail is 0.

The control variables are from the National Board of Education application data from 1998-2012. This data includes basically every upper secondary school application made during this time period. There was some variation over time in the data included in the database. However, data for each year included the national identification number of each applicant and the identification number of the comprehensive school from which the applicant had graduated. Also, the grade point average on which selection into upper secondary school is based was included for each applicant. There have been some minor changes in this variable over the years and we tested for breakpoints but did not find any. More detailed information on grades was also included. This was not used in the main analyses but were exploited in some auxiliary calculations.

A student may appear in the application data multiple times if s/he applied to secondary education multiple times. This poses no problems if the data for each time is consistent. For some students this was not the case. For these the latest information

Table 1: Descriptive statistics

	Mean	SD	Obs
Matriculation avg.	4.22	1.06	363176
Comprehensive avg.	8.41	0.73	363176
Swedish-speaking	0.06		363176
Non-Finnish/Swedish speaking	0.01		363176
Female	0.58		363176
Comprehensive math.	8.31	1.07	362817
Comprehensive Fin/Swe	8.45	0.85	362721
Matriculating students	74.38		4883

was used. If this was not enough to produce a unique value for all variables, the row containing most information was chosen.

For some years, the application data contained the address of the applicants. These addresses were geocoded and the coordinates of these students were mapped into a grid square in the Grid Database of Statistics Finland. The grid database contains information on $250m \times 250m$ grid squares in a grid covering the whole country. The data are averages of socioeconomic and demographic variables in each grid square. As addresses were available for only a small subset of students, grid data was not employed in the main analyses, but used in some auxiliary calculations. The grid variables used were: unemployment rate, share of owner-occupied housing, population, share of highly educated and median income.

The matriculation data and application data were merged using the national identification number of students. As there is usually at least a three-year lag between application to an upper secondary school and matriculation, merging was started at the year 2002. From the original matriculation database data on 397 500 matriculations was recovered. The merged data contains information on 386 000 matriculations.

After this, some additional data was discarded. First, some upper secondary schools operate as night schools or so-called continuing education schools. These were dropped out of the data, as their operations and goals differ markedly from those of "ordinary" USS's. Also, there have been a number of school closings and mergers during the time period. Closed schools were left in the data. The precursors of merged schools were treated as parts of a merged entity during the whole period. Significant effort was expended to observe as many of the mergers as possible. This was complicated by the fact that there is no comprehensive database on school closings and mergers. A list of schools operating at each year exists only on paper format for the first years in our period. Finally, students with information lacking on vital variables were dropped. In the final data there are around 360 000 matriculating students. Of these, about 90 000 could be connected to the grid database. In Table 1 some descriptive statistics can be found.

6 Estimation results

6.1 Estimation

Three kinds of value-added models were estimated, each based on the methods described in Section4. The first value-addeds are based on all the data. These value-addeds are the main results, and the phrase value-added estimates without additional qualifications refers to these. The second type of value-addeds are the so-called jackknife value-addeds, which estimate each year's value-added based on data on all other years. These are used in various predictive exercises. Finally, for comparison, we calculate so-called Kane-Staiger type estimates, which differ from the first two in that they assume constant school effects over time.

As described in Section 4, first a residualized matriculation score was calculated for each student. In the main model the residualization was based on a model in which the exam average was regressed on (upper secondary) school fixed effects, a third degree polynomial in comprehensive school grade average, mother tongue (Finnish, Swedish, Other), gender, year dummies and finally a comprehensive school variable. The comprehensive school variable was included for a number of reasons. First, there is some evidence that the grade scales differ somewhat between comprehensive schools (see[24]). If this is true, the comprehensive school grade average is by itself not enough to control student quality. Another reason to include a comprehensive school effect is an attempt to limit selection bias. As discussed in Section 5, there is very limited information on student characteristics, such as socioeconomic status. Thus there is significant scope for selection problems if no additional controls are included. Including a comprehensive school effect may limit this problem, assuming that part of the selection has already happened in comprehensive school, for example because of parents' location choice.

The student-level residualized scores were then aggregated into school-year-level means. Using each school that had observations with at least k years between them, the autocovariances $\sigma_{A,k} = Cov(\overline{A}_{j,t}, \overline{A}_{j,t-k})$ were estimated. In these calculations weights corresponding to numbers of students in both time periods combined were used.

Figure 1 presents the autocorrelation function calculated using these convariances. The numbers on which the figure is based are given in Table 2

Conditional of all assumptions, the autocorrelation function has an intuitive interpretation. Think of a situation in which you are asked to predict the school effect of a school k years into the future based on this years results. The autocorrelation coefficient is the coefficient with which the current effect must be multiplied to get the best prediction of the future score. For example, the one-period autocorrelation of 0.63 means that next year's predicted effect is 0.63 times the curren effect. So if the school is for example one standard deviation worse than the average school this year, it is predicted that next year it is 0.63 standard deviations worse. A similar prediction for 9 years in the fututre is that the school is only a third of a standard deviation worse than average.

The estimated autocorrelation function indicates, that the estimated school effects have a tendency to revert in the direction of the mean, but that there is a persistent component, as the autocorrelation is not eliminated completely even in a decade.

Table 2: Autocovariance and autocorrelation

	Autocovariance	Autocorrelation
Lag 1	0.021	0.628
Lag 2	0.018	0.550
Lag 3	0.016	0.489
Lag 4	0.015	0.469
Lag 5	0.014	0.419
Lag 6	0.013	0.401
Lag 7	0.012	0.352
Lag 8	0.012	0.366
Lag 9	0.012	0.341

Figure 1: Autocorrelation

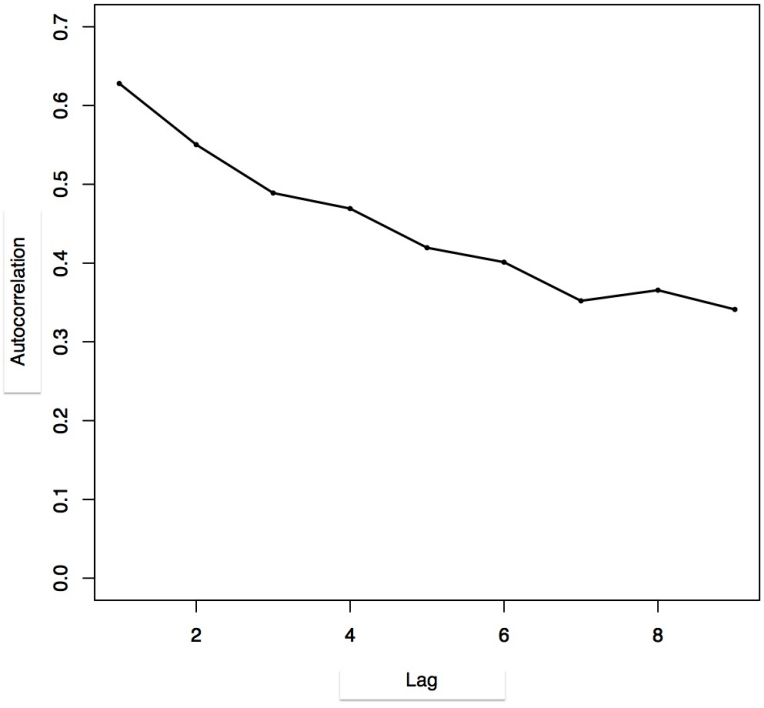


Table 3: Standard deviation of the school effect.

	SD	1. lag	Quadr.
Total SD	0.76		
Within-School SD	0.74		
Between-School SD	0.16	0.14	0.15

An interesting question is how much of the total variation in residualized scores is attributable to within-school and how much to between-school variation. A natural measure of between-school variation would be the standard deviation of the school effect. When all assumptions are satisfied, an estimate of the school effect variance is given by $\hat{\sigma}^2_\mu = \hat{\sigma}^2_A - \hat{\sigma}^2_\varepsilon$, the square root of which is an estimate to the standard deviation. To see whether this figure is plausible, it is possible to quantify the standard deviation also by other means. For example, the covariance of the school effects for two adjacent years is always lower than the variance, so a lower bound for the standard deviation is given by the square root of the one-lag autcovariance. It is also possible to try to extrapolate the standard deviation using the autocovariance function [4]. The results of these calculations are given in Table3. Naturally, if there is bias in our value-added estimates, these figures will also be biased.

It is seen that the total standard deviation of residualized scores is about 0.76 grade points. Individual-level SD or within-school SD is 0.74 and estimates for school effect SD are around 0.15. This means that only about 5 % of total residualized score variation is due to schools and 95 % due to individuals[5].

When the covariances have been calculated it is possible to use them as described in Section 4 to calculate the value-added estimates. In these calculations covariances were calculated up to 9 lags. The rest of the autocovariances were set to the same level as the 9th lag, because autocovariances for higher lags would have been based on too few data points. For the main estimates the value-added for each year is the linear prediction of a year's school effect given all residualized scores, including the ones for the year in question. This means that both past and future scores were used in the prediction. Value-addeds were estimated for years 2002-2013 for school-year combinations with at least two students.

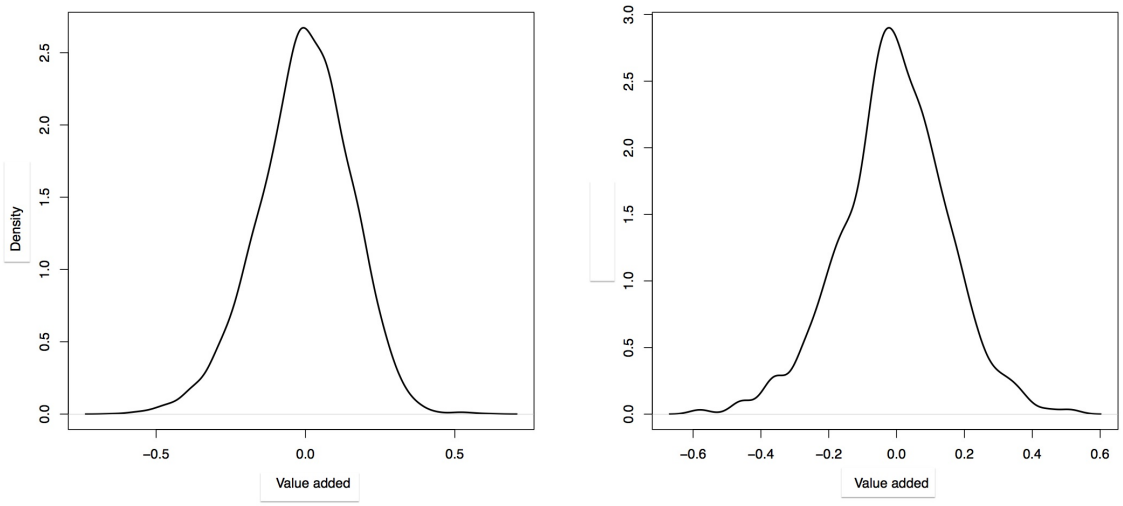
For the purposes of some predictive excercises discussed below we also estimated jackknife value-addeds, with prediction based on information on all other years except the one in question. Jackknife value-addeds were calculated for years 2002-2014. Finally, for comparison purposes we calculated the so-called Kane-Staiger type estimates with constant school effects.

The main estimates and the jackknife estimates agreed quite well with each other. There were some differences due to the different information used in the estimation.

[4] A quadratic model regressing the autocovariance function on time was fitted. The intercept of the model is the estimate for the variance

[5] These shares are based on variances. The figure 5 % is the so-called intra class correlation and it is calculated by dividing the square of the school effect SD by the square of total SD.

Figure 2: Distribution of value-added estimates



(a) Value-added estimates 2002-2013

(b) Value-added estimates 2013

Correlation between the two value-added estimates was over 0.9. Also, the jackknife estimates predicted the main estimates well, and there was no evidence of systematic differences between the two.

As one possible use of value-added estimation is ranking of schools, we also calculated rank correlations between the two estimates. Average yearly Spearmann correlations were also over 0.9, but average Kendall correlation was only 0.7. This demonstrates the sensitivity of rankings to even small changes in value-added estimates, a topic to which we will return below.

6.2 Distribution of value-addeds

The distribution of the value-added estimates is described in the density plots of Fig. 2. Panel 3a plots the density of all value-addeds in the years 2002-2013, while panel 3b plots the density for year 2013 alone. Yearly descriptive statistics are given in Table 4. Recall that value-added for the average school is normalized to zero.

It is clear that there are differences between schools. The difference between the top and bottom schools is around one grade point each year. If the estimates are unbiased this means that a the matriculation exam average for a randomly chosen student would be one grade point higher in the top than in the bottom school. This is a significant difference, as it means that the student would score on average one grade higher in all four subjects she / he is takin the test in. Such a difference would also significantly enhance the student's chances of getting into a university.

However, the differences are this large only between top and bottom institutions. Look-

Table 4: Descriptive statistics for estimated value-addeds

	2002	2003	2004	2005	2006	2007
Min	-0.50	-0.48	-0.50	-0.47	-0.52	-0.56
1st Decile	-0.24	-0.21	-0.19	-0.18	-0.21	-0.18
1st Quartile	-0.12	-0.12	-0.09	-0.10	-0.10	-0.09
3rd Quartile	0.08	0.11	0.13	0.10	0.09	0.09
9th Decile	0.17	0.19	0.20	0.20	0.16	0.18
Max	0.37	0.43	0.38	0.39	0.44	0.51
	2008	2009	2010	2011	2012	2013
Min	-0.65	-0.55	-0.59	-0.54	-0.56	-0.58
1st Decile	-0.19	-0.19	-0.20	-0.22	-0.22	-0.20
1st Quartile	-0.08	-0.08	-0.10	-0.11	-0.10	-0.10
3rd Quartile.	0.11	0.11	0.09	0.10	0.09	0.09
9th Decile	0.19	0.21	0.20	0.19	0.17	0.18
Max	0.62	0.56	0.55	0.50	0.53	0.51

ing at the results more closely, one can see that most schools are much closer to each other. Each year the mid-dle 80 % of schools are within at most 0.4 grade points of each other. The similar figure for the middle 50 % of schools is only 0.2 grade points.

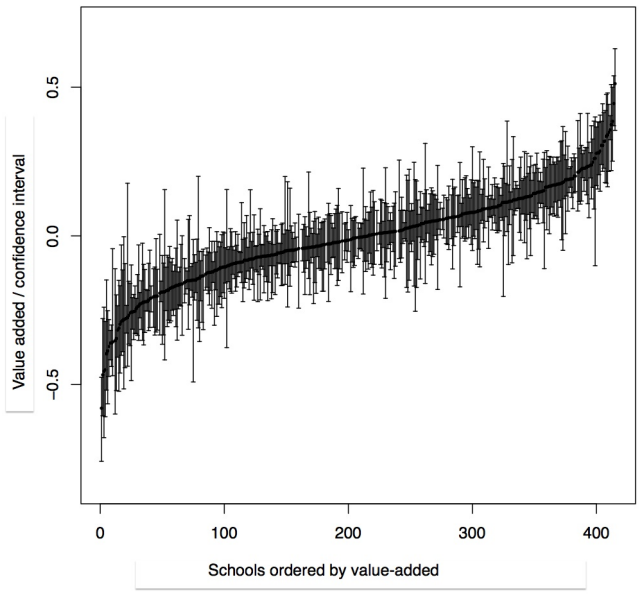
These results lead to at least two questions. One concerns the statistical significance of the results. As most schools are close to each other, it is natural to consider whether these small differences are due to estimation error and chance instead of systematic differences. The other, related question concerns the persistence of value-addeds in time. Do the same schools inhabit the top and bottom of the distribution each year, or is there constant change in the extremes of the distribution. These questions cannot be answered by looking at Fig. 2 and Table 4.

6.3 Quality differences and rankings of schools

Start with the first question. Could the differences between schools be due to chance and estimation error? Typically one would approach this question by performing appropriate statistical tests and calculating confidence sets. Doing this is not at all straightforward with hundreds of value-addeds from multiple years. This is why we choose an informal approach. We calculate a 95 % confidence interval for each value-added estimate using a block bootstrap. We take 1000 samples with replacement from the school population. Each of the samples has the same number of schools as in the original data. We then calculate value addeds based on all these samples and take the appropriate quantiles as the bounds of the confidence intervals.

The results are presented in Fig. 3. On the horizontal axis are schools ordered by value-added. Value-addeds and confidence intervals for the year 2013 are on the vertical axis. It is immediately obvious that many of the confidence intervals are quite broad. There is considerable uncertainty concerning the value-addeds. For many schools the

Figure 3: Value-addeds with confidence intevals, year 2013



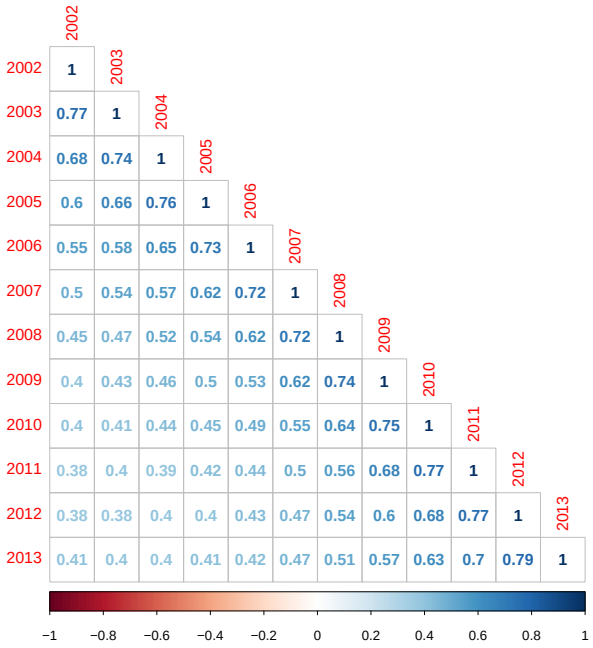
lower bound of the confidence interval is below zero and the upper one above zero. Such schools do not significantly differ from average. In the year 2013 plotted in the figure, the schools indistinguishable from the average make up 46 % of all schools. In comparison, 27 % of schools were better than average while a similar percentage were worse.

It can also be seen from the figure that there is considerable overlap for the confidence intervals, even for schools quite distant from each other in ranks. Looking for overlapping confidence intervals is of course not an entirely correct way of distinguishing value-addeds statistically. However, the statistical properties of rankings are far from simple (see. [8, 9]) and that is why we adopt this informal approach of looking at value-addeds. It seems clear from the figure that most schools in the middle of the value-added distribution are statistically indistinguishable, and thus any ranking of them based on value-addeds must be meaningless. Only between the very top and very bottom of the distribution can be ranked according to value-added.

6.4 Persistence

It is not possible to rank schools except the very best and very worst schools. But are these top and bottom schools the same all the time or is there significant turnover at the extremes of the distribution? One answer to this question was already seen in subsection 6.1, in which autcorrelation of the mean residualized scores was considered. The result was that there is a persistent component in these means, but also significant time series variation. Let us now make this more concrete.

Figure 4: Rank correlation of value addeds (Kendall)



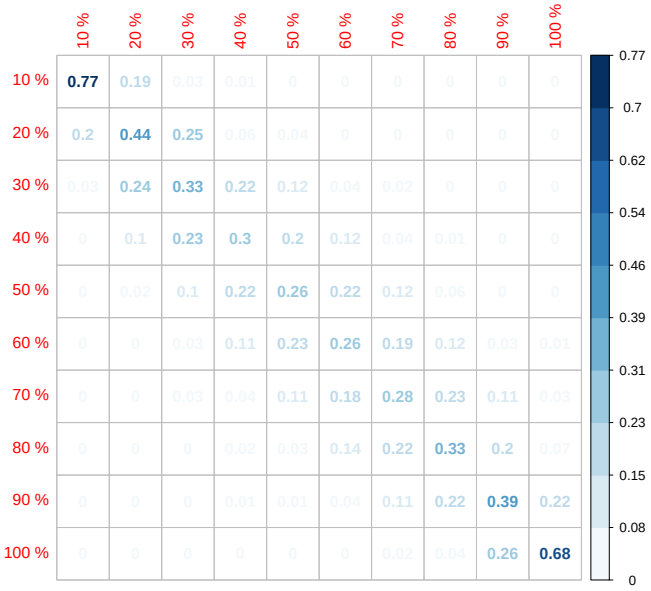
One straightforward way of looking at persistence is calculating the rank correlation between value-addeds of different years. The results of such a calculation is given in Fig. 4.

The figure gives a correlation matrix based on Kendall's rank correlation for value-added estimates. It can be seen that the correlation matrix reinforces earlier results. The rankings based on value-addeds for adjacent years are similar, but there is also persistence after many years. This persistence is however far from perfect, and it would be an exaggeration to say that rankings based on value-added would be stable over time.

Another way of looking at persistence is to look at transitions between adjacent years. These may be described using an estimated transition matrix. Such a matrix is presented in Fig. 5. Each rows corresponds to a starting decile in the value-added distribution and each column corresponds to the decile next year. The number in the appropriate cell is the estimated probablitity of transitioning from the starting decile to the target decile. For example, the mnumber 0.25 on the second row and in the third column means that a school in the second worst decile has a probability of 0.25 of transitioning into the next decile.

The transition matrix shows that almost 80 % of the bottom 10 % of schools stay in the bottom in the next year. This is a rather high number, but hardly close to one. Almost a fifth of schools at the bottom make it to the next decile the following year. The probability of staying at the top is similar. Most schools at the top stay at the top also next year, but a fourth drop down one decile, and a similar number of schools rise to the top from the next highest decile. Persistence is thus strong for adjacent years, but recall

Figure 5: The decile transition matrix



that during our period of interest there are over 10 transitions, which will create quite a lot of turnover during the period.

Fig. 6 presents another transition matrix.

This gives transition probabilities between the top and bottom percentiles and the middle 98 %. The probability of staying in the top percentile is 60 % and conversely a probability of 40 % of dropping into the middle. The probability of making it out of the bottom percentile is 35 %.

Still another way of looking at persistence is Fig. 7.

The figure plots value-addeds and rankings between adjacent years. It reinforces the previous results: there is high correlation between adjacent value-addeds, and significant but far lesser correlation between ranks. This means that any rankings based on value addeds will be unstable over time for all schools in the middle of the value-added distribution, that is, for most schools.

7 Robustness and bias

7.1 Prediction properties

In this section we attempt to assess the reliability and robustness of our results. First we look at the out-of-sample predictive ability of value-addeds and then at forecast bias. In Section4 the assumptions underlying the method were spelled out. There were two sets of assumptions. The first assumptions guarantee that the method produces an estimate

Figure 6: Lukioiden transitiomatriisi, hännät

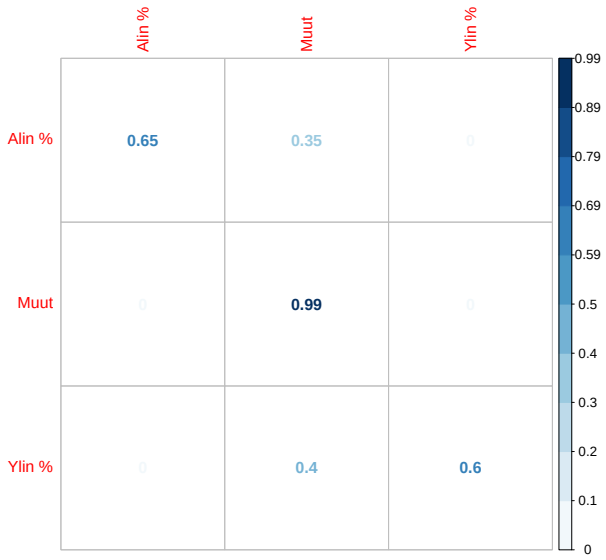
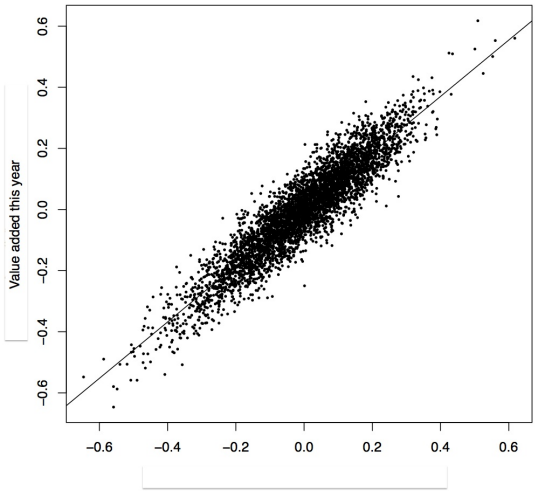
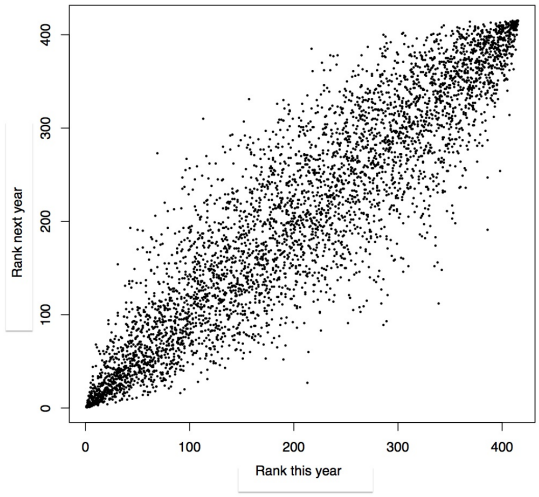


Figure 7: Adjacent years



(a) Adjacent value-addeds

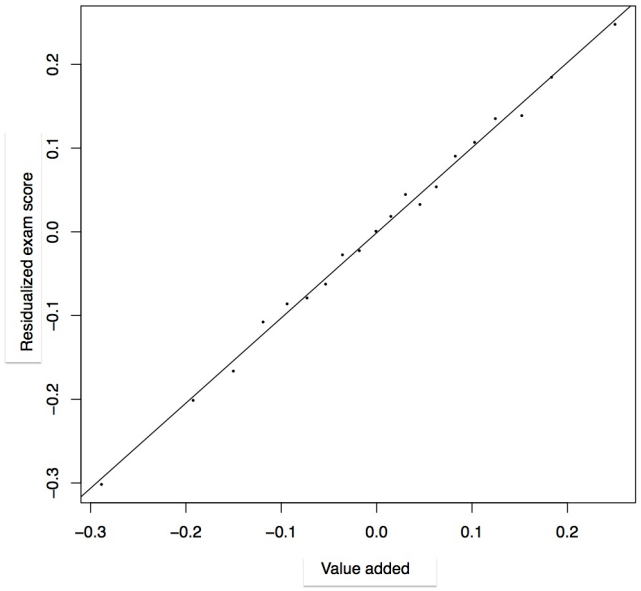


(b) Adjacent ranks (N.B. a higher rank is a better rank)

Table 5: Prediction model results

	Estimate	Std. error	t-value	p-value
Constant	-0.001	0.007	-0.177	0.860
Value added	1.017	0.022	45.751	0.000

Figure 8: Value added and residualized scores



of the best linear prediction for residualized matriculation exam scores. If these first assumptions hold, value-added estimates should on average predict residualized scores. As noted in Section 4 the assessment of predictive ability must be based on the jackknife value-added estimates. If value addeds are indeed linear predictions of residualized scores, then a prediction model in which residualized scores are predicted using the jackknife value addeds should produce a coefficient of unity for the value-added.

Estimation results for such a model are given in Table 5.

The estimated coefficient for the value-added measure is 1.017 and its standard error[6] is 0.022. The coefficient does not significantly differ from unity. Therefore there is no evidence against the hypothesis that the value addeds are indeed linear predictions[7]. This has been depicted graphically in Fig. 8. The figure is a scatterplot of value addeds and residualized scores. To make the figure readable, the observations have been classified before plotting. Each point is the mean of one of twenty value added classes. The

[6]The standard error estimate is calculated with two-way clustering with respect to school and year.
[7]Also the combined hypothesis of zero intercept and unit coefficient is not rejected.

value addeds have been assigned into these classes in rank order. On the horizontal axis the mean of each value-added class is depicted while the vertical axis measures the corresponding mean of residualized scores. The points are close to a 45° line through the origin. The conclusion is therefore that the prediction properties of the value-added measures are as they should be.

7.2 Bias calculations

As already mentioned many times, selection can invalidate our value-added measures as estimates of the true school effect. True school effects could be easily estimated if students were assigned randomly to schools, but in reality the coupling of schools and students is decidedly non-random. As discussed in Subsection 4.6 quantifying the amount of bias is a difficult problem. Something can, however, be done. In the subsection a method of quantifying forecast bias was explained. Value addeds are forecast unbiased if they would predict on average correctly differences in exam success in if students were randomly assigned.

The basic idea was to use additional control variables to observe forecast bias. The new variables are to be chosen to be as strongly correlated as possible with the possible causes of selection bias. For example, if study motivation or parental investment are seen to be the main potential causes of bias, then the new variables could be socioeconomic characteristics of the students or their parents. As far as socioeconomic characteristics predict motivation and parental investment, the inclusion of these variables should decrease bias in the estimates. Therefore, if inclusion of new variables changes results significantly, this can be interpreted as evidence of bias. If results are not significantly affected by inclusion of new controls, there is no evidence of bias. The details of the method are explained in Subsection 4.6.

First, we assess the impact of including the student's comprehensive school in the controls. As discussed in Section 6 the reason for controlling for the comprehensive school was on the one hand to take into account possible differences in grading practices between comprehensive schools and on the other to diminish the scope for selection problems. The comprehensive school is therefore an important control and it is interesting to see how much impact it has on the results. That is why we went through the forecast bias testing procedure to test for the significance of the comprehensive school variable. We took as starting point a model with otherwise the same controls as our basic model but without the comprehensive school variables. Then comprehensive school effects were introduced as "new" variables, and the bias of leaving the comprehensive school effects out was estimated using the bias procedure.

The result is on the first row of Table 6. Estimate for forecast bias caused by leaving out the comprehensive school effects is 60 %. This means that comprehensive school effects are very important. According to this result, leaving these effects out will cause differences between upper secondary schools to be grossly overestimated.

Next we tested for bias using socioeconomic variables. As we had no access to student-level socioeconomic information we used data from the Grid Database of Statistics Finland. In the database Finland is divided into $250m \times 250m$ grid squares with information

Table 6: Calculations of bias

	Left-out / new variables	Estimate of bias
1	Comprehensive school effects	60.3 %
2	Math / Fin / Swe grades	0.1 %
3	Socioeconomic	0.7 %

on average socioeconomic variables in each grid square. We had address data for about 90 000 students in our database and used the grid square averages as proxy socioeconomic variables for these students. Two slightly different approaches were used, both produced very similar results[8]. The results are given on the third row of Table 6. Estimated forecast bias is under 1 %, which means that selection according to anything correlated with socioeconomic variables does not seem to be a big problem. It must, however be remembered that no student-level information was available and even grid square -level information was available to only a small subsample.

Finally, we tested could the model be improved by controlling for additional student quality variables, in this case math and Finnish / Swedish grades. We added third degree polynomials in each and tested for forecast bias. The results are on the second row of Table 6 and show that no evidence of significant bias was found. In conclusion, no evidence was found of significant bias caused by selection on anything correlated with socioeconomic variables. Also, no evidence on bias caused by imperfect contolling for comprehensive schools was found. However, omitting comprehensive school effects will cause significant overestimation of upper secondary school effects.

8 Discussion

According to our results, there are significant differences between Finnish upper secondary schools. The difference between the top and the bottom schools is about one grade point each year. The exam score variable used in our analyses is the average score of four matriculation exams, each graded on a scale of 0 to 7. This means that a randomly selected student assigned to the top school would have received an average score one point higher than in the bottom school. This of course means that on average the student would have scored one point more in all four of the exams included in the average. Given the Finnish system, where matriculation exam results determine in part for example university entry, such a difference is highly significant.

The results should be interpreted with caution, however. These large differences are found only between the very top and the very bottom. Our value-added measures for most schools are very close to each other. In fact they are so close, that distinguishing between them in a statistically meaningful way is impossible. Also, while there is significant

[8]The first approach used only the observations with address data. In the second approach all observations were used and missing data was coded as 0 and a missing information dummy was included in the controls.

Table 7: Top / bottom 5 % for the whole period

	Always top 5 %	Always bottom 5 %
1	Mynämäen lukio	Närpes gymnasium
2		Vörå samgymnasium

persistence in value-added measures, during the period of about a decade that we observe, value addeds can experience significant change over time.

These caveats apply even more forcefully to any rankings or league tables based on the value-addeds. While rankings are strongly correlated over time, there is significant instability in rankings. Most of the schools cannot be meaningfully ranked.

The possibility of selection bias is also cause for caution. Even though the methods used are sophisticated, it is still unclear how much bias is caused by selection. We have tried to find evidence of selection bias, but failed to find any. This does not mean that there is none.

One result is that the highly selective "elite" schools in bigger cities, notably in the Helsinki metropolitan area that dominate league tables based on raw exam success, are all found in the undistinguished middle of the value-added distribution. Their success in exams is outstanding, but they also attract the most succesful students from the best comprehensive schools. Of course, it must be borne in mind that value added or exam success in general is not the only reason to attend elite institutions. These schools can tie a student to a network which can prove valuable for example on the job or marriage markets.

One of the results is that any rankings based on value-added measures are quite unstable. This presents the question whether there are any schools that are consistently very good or bad. One way of looking at this is to see which institutions have remained at the top of the distribution during the whole period of observation. These schools (perhaps the names are more interesting to Finns than readers of other nationalites) are listed in Tables 7 and 8. In each the schools are in no particular order. Table 7 lists all schools that have remained in the top or the bottom 5 % for the whole period of observation. As can be seen from the table, there are extremely few of these. Only one school has remained in top 5 % for the whole period 2002-2013. Two schoos have remained in the bottom 5 %.

The period of observation is quite long, and the first years can seem like ancient history. Table8 lists the schools that have been in the top or bottom 5 % in and after 2008. There are relatively few schools in this list as well. Slightly more schools have remained in the bottom than in the top. This is because as we saw above, persistence in the bottom is somewhat higher than in the top.

Table 8: Top / bottom 5 % from 2008

	Top 5 % from 2008	Bottom 5 % from 2008
1	Leppävirran lukio	Jakobstads gymnasium
2	Mynämäen lukio	Langinkosken lukio
3	Reisjärven lukio	Kristinestads gymnasium
4		Närpes gymnasium
5		Kronoby gymnasium
6		Merikarvian lukio
7		Konneveden lukio
8		Rautjärven lukio
9		Topeliusgymnasiet i Nykarleby
10		Pälkäneen lukio
11		Korsholms gymnasium
12		Vörå samgymnasium

None of the top schools can be called elite institutions. They are not in big cities but in small towns. These schools do not accept only students with top grades from comprehensive schools, but instead take in a broad selection of students in their area. It is at the moment unclear to the authors why the value-added results are so good for these schools. The bottom institutions are also provincial. But here the most interesting thing is the overrepresentation of Swedish-speaking minority schools. We are lookin at this phenomenon, but at the moment it remains a mystery.

References

[1] Aamulehti. Stt:n lukiovertailu: Viime vuoden ykkönen romahti sijalle 56.

[2] Derek C Briggs and Jonathan P Weeks. The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, 36(5):616–637, 2011.

[3] Gary E Chamberlain. Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, page 201315746, 2013.

[4] Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic review*, Forthcoming.

[5] Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic review*, Forthcoming.

[6] Dan Goldhaber and Duncan Chaplin. Assessing the rothstein test: Does it really show teacher value-added models are biased? Technical report, Mathematica Policy Research, 2012.

[7] Dan Goldhaber and Michael Hansen. Is it just a bad class? assessing the long-term stability of estimated teacher performance. *Economica*, 80(319):589–612, 2013.

[8] Harvey Goldstein and Michael JR Healy. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 175–177, 1995.

[9] Harvey Goldstein and David J Spiegelhalter. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 385–443, 1996.

[10] C Kirabo Jackson, Jonah E Rockoff, and Douglas O Staiger. Teacher effects and teacher-related policies. *Annual Review of Economics*, (0), 2014.

[11] Brian A Jacob, Lars Lefgren, and David P Sims. The persistence of teacher-induced learning. *Journal of Human resources*, 45(4):915–943, 2010.

[12] Thomas J Kane, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger. Have we identified effective teachers? validating measures of effective teaching using random assignment. *Seattle, WA: Bill and Melinda Gates Foundation*, 2013.

[13] Thomas J Kane and Douglas O Staiger. Improving school accountability measures. Technical report, National Bureau of Economic Research, 2001.

[14] Thomas J Kane and Douglas O Staiger. The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives*, 16(4):91–114, 2002.

[15] Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research, 2008.

[16] HoonHo Kim and Diane Lalancette. Literature review on the value-added measurement in higher education. Technical report, OECD, 2013.

[17] Josh Kinsler. Assessing rothstein's critique of teacher value-added models. *Quantitative Economics*, 3(2):333–362, 2012.

[18] Tanja Kirjavainen. *Kirjavainen, T. Essays on the Efficiency of Schools and Student Achievement*. Number 53 in VATT Publications. VATT, 2009.

[19] Cory Koedel and Julian R Betts. Does student sorting invalidate value-added models of teacher effectiveness? an extended analysis of the rothstein critique. *Education finance and policy*, 6(1):18–42, 2011.

[20] Jorma Kuusela. *Lukioiden tuloksiin vaikuttavista tekijöistä.* Opetushallitus, 2003.

[21] Daniel F McCaffrey, JR Lockwood, Daniel Koretz, Thomas A Louis, and Laura Hamilton. Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, 29(1):67–101, 2004.

[22] Daniel F McCaffrey, Tim R Sass, JR Lockwood, and Kata Mihaly. The intertemporal variability of teacher effect estimates. *Education finance and policy*, 4(4):572–606, 2009.

[23] Carl N Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

[24] Najat Ouakrim-Soinio. *Toimivatko päättöarvioinnin kriteerit?* Raportit ja selvitykset. University of Helsinki, Faculty of Behavioural Sciences, 2013.

[25] Jenni Pääkkönen and Laura Ansala. Kouluvaikutus ja tuloksellisuusrahoitus lukiokoulutuksessa. Technical report, VATT, 2013.

[26] Jesse Rothstein. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, 4(4):537–571, 2009.

[27] Jesse Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

[28] Helsingin Sanomat. Hs:n lukiovertailussa kaikki maan koulut: Etelä-tapiolan lukio ykkönen, 2014.