CPB Netherlands Bureau for Economic Policy Analysis

CPB Discussion Paper | 230

Teacher evaluations and pupil achievement

Evidence from classroom observations

Marc van der Steeg
Sander Gerritsen

Teacher evaluations and pupil achievement:

Evidence from classroom observations[1]

Marc van der Steeg*
CPB, Netherlands Bureau for Economic Policy Analysis
m.w.van.der.steeg@cpb.nl

Sander Gerritsen
CPB, Netherlands Bureau for Economic Policy Analysis
s.b.gerritsen@cpb.nl

Abstract

This paper investigates the relationship between teacher evaluations, conducted by trained evaluators, and pupil performance in primary education in a large city in the Netherlands. Teacher evaluations are based on a detailed rubric containing 75 classroom practices considered to be crucial for effective teaching. We obtain a set of estimates that suggests that the score on this rubric significantly predicts pupil performance gains. Estimated test score gains are in the order of 0.4 standard deviations in math and grammar if a pupil is assigned to a teacher from the top quartile instead of the bottom quartile of the distribution of the evaluation rubric. These are relatively large differences in pupil outcomes, suggesting that evaluations based on the rubric measure teacher practices that matter for pupil performance. This suggests that the rubric seems to have potential for teacher evaluations and teacher effort.

JEL Codes: I2
Keywords: teacher evaluation, pupil performance

*Corresponding author: M.W.van.der.Steeg@cpb.nl

[1] The authors like to thank Debby Lanser, Trudie Schils, Dinand Webbink and Bas ter Weel for valuable comments on earlier versions of this paper.

1. Introduction

Research on the impact of teacher quality on student achievement consistently shows that teachers matter for student achievement. Children assigned to a teacher with a one standard deviation higher quality gain in terms of achievement in the order of 0.10 to 0.25 standard deviations.[2] In addition, the economic returns to higher quality teachers can be substantial. For example, Chetty et al. (2011) show that children assigned to teachers with a higher 'value-added' (i.e., teachers that produce larger achievement gains) attend college more often, earn more and live in better neighborhoods. Staiger & Rockoff (2010) predict a total gain of 330 to 760 thousand dollar in lifetime income for a class that has a one standard deviation better qualified teacher.

It is less clear which teacher characteristics or practices matter. Traditional observable characteristics of teachers - often used to determine teacher pay levels - have only little predictive power for measuring differences in teacher quality. With respect to teacher qualifications most studies do not find a relationship between the teacher's highest attained education level and teacher quality.[3] With respect to work experience, most studies show that teachers gain in terms of effectiveness in the first two or three years of their career, but that this experience effect levels off after this period.[4]

A lack of knowledge about effective teacher characteristics and practices is problematic for policymakers and school leaders that aim to improve and reward teacher quality. Recent research in the United States reveals that teacher ratings or evaluations made by school principals, mentor teachers or trained evaluators have predictive power for student achievement.[5] Estimates from these studies show that, depending on the domain (reading or math) and type of evaluation, a one standard deviation higher evaluation score is related to 0.05 to 0.14 of a standard deviation higher student achievement scores.

In this paper, we use teacher evaluations to estimate the relationship between a teacher's evaluation score and pupil achievement. The detailed evaluations were carried out by trained evaluators in seven elementary schools in a large city in the Netherlands. The level of detail sets our study apart from previous studies. Our measure includes a rubric of 75 associated teacher practices and 18 standards. This is more than double the number of standards and associated practice descriptions relative to the studies conducted previously. For example, the Cincinatti's Teacher Evaluation System (TES) studied in Kane et al. (2011) and Tyler et al. (2010) has 8 standards and 29 associated practice descriptions. Another difference is that Cincinatti's TES has 4 scoring levels for each particular practice (from unsatisfactory to proficient), whereas our rubric has 2. After the observation of the teacher was made, the evaluator had to indicate whether or not the practice was shown by the teacher.

[2] See e.g., Rockoff (2004), Rivkin et al. (2005), Aaronson et al. (2007), Kane & Staiger (2008) and Hanushek & Rivkin (2010).
[3] See Hanushek & Rivkin (2006) for a review study.
[4] See, among others, Rivkin et al. (2005), Clothfelter et al. (2006) and Jacob (2007), and Staiger & Rockoff (2010).
[5] See Jacob & Lefgren (2008), Rockoff & Speroni (2010), Tyler et al. (2010) and Kane et al. (2011).

Our main findings can be summarized as follows. We obtain estimated coefficients that suggest that a higher score on the rubric is statistically significant related to higher pupil achievement. A one standard deviation higher score on this rubric is associated with a 0.15 standard deviation higher score in math, a 0.18 standard deviation higher score in grammar and a 0.11 standard deviation higher score in reading. Gains in pupil achievement are relatively large if a teacher from the bottom quartile of the score on the rubric is replaced by a teacher from the top quartile. Estimated gains range from 0.24 (reading) to 0.44 (grammar) standard deviations. These gains are considerably larger than the ones found in Kane et al. (2011) for Cincinatti's TES. We also observe that the rubric seems to be particularly capable of identifying the weakest teachers, but seems less capable of differentiating between an average teacher and an excellent one.

Our contribution to the literature is twofold. First, we corroborate results that have been found in previous studies in the United States. In particular, differences in teacher quality are large and teacher evaluations through classroom observations are useful in identifying such differences. Second, the rubric used in this paper seems to do a somewhat better job in identifying differences in teacher quality compared to rubrics assessed in earlier literature, especially in identifying the weakest teachers. This could be due to the higher level of detail (i.e., more teacher practices) or to different competences being assessed by our rubric. Our rubric seems particularly more capable of identifying the worst teachers, which is important since weak teachers have a negative impact on a student's achievement and later socioeconomic outcomes. This study, together with the small but growing literature that has recently emerged, suggests that evaluations made by trained experts have potential in identifying heterogeneity in teacher quality. The results of this research could be used to address problems of low teacher quality and to provide and design effective incentive schemes to improve teacher quality.

We proceed as follows. Section 2 describes the evaluation rubric and the data. Section 3 presents the empirical strategy. Section 4 shows the estimation results. Section 5 concludes.

2. Data

We use data on pupils and teachers from grade 1 to 8 from seven elementary schools from a school district in a large city in the Netherlands. Elementary education in the Netherlands starts when children reach the age of 4 and ends when they are 12 years old. School age starts at age 5, but the vast majority (>95 percent) of children enters at age 4. The seven schools participated in a teacher evaluation project that was launched by the municipality. The pupil data contain information on math, grammar and reading test scores from the end of the school year 2011/2012 and previous test scores from the end of the school year 2010/2011.[6] Besides information about test scores, we have obtained detailed information about pupil background

[6] Pupil test scores are from tests that are developed by the national test developing agency CITO. The majority of primary schools in the Netherlands use these tests to monitor progress of their pupils.

characteristics, such as age, gender, educational level of the parents and whether the child lives in a one parent family. The teacher data contain teacher experience and the scores on the rubric with the 75 classroom practices. Professionals have identified this list to reflect good teacher practices. Next to the data on teachers, we have obtained classroom information, such as class size, the fraction of girls and the fraction on pupils whose parents are low educated.

In the empirical analysis we use standardized test scores for math, grammar and reading from the school year 2011/2012 as dependent variables. Test scores have been standardized by school year and grade.

Our independent variable is the total score on the teacher evaluation system (TES). A detailed rubric has been constructed by educational professionals for the purpose of citywide monitoring of teacher quality. The rubric consists of 18 standards and 75 associated classroom practices that are expected to reflect good teacher practices.[7] These classroom practices are defined in three domains: pedagogical competence, didactical competence and organizational competence. The Cronbach's Alphas for all 75 items of the rubric and for the items in the three domains respectively are 0.96, 0.85 (15 items), 0.94 (46 items) and 0.84 (14 items). They are all larger than 0.8, suggesting that the internal consistency or construct validity of the rubric is sound.

Here we only discuss the most salient details of the rubric; appendix table A.1 provides an overview of the 18 standards of the rubric. While teaching a class, teachers were scored on this rubric by professional evaluators that have been specifically trained for the job. The training of the evaluators had a particular focus on consistency, in order to prevent different evaluators to evaluate differently as much as possible. The evaluations were announced evaluations and each evaluation was done by one single evaluator. The teachers had to give a class in which they could show all 75 classroom practices. Teachers could either show a classroom practice or not, with the evaluator denoting a 1 if the teacher showed the competence and 0 if not. Hence, the score on the rubric may range from 0 to 75.

Teachers were evaluated twice, once in the beginning of the school year and once at the end of the school year. Following Kane & Staiger (2012), we take the average of the start and the end score on the rubric. They advise to use multiple classroom observations for each teacher to obtain a more reliable picture of the true quality of the teacher. For 106 teachers we have both scores on the TES. For 19 teachers the end score is missing. There are multiple reasons for the missing values on the end score: some teachers left school during the school year 2011/2012, other teachers were not present in the week the evaluations were carried out due to illness or pregnancy, and yet others were in the middle of a dismissal procedure. These 19 teachers were relatively weak teachers as their score on the rubric was below average (i.e., on average 10 points lower). For these teachers we imputed the score from the end of the school year with the score from the start of the school year, and included an indicator for missing

[7] The official competence requirements for teachers that are used by the Education Inspectorate of the Netherlands and that are part of the Law on Occupations in Education (*Wet Beroepen in Onderwijs*) have been transferred to corresponding observable classroom practices in the rubric.

end score in our models. To investigate whether or not our results are influenced by these missing observations, we will present estimates for both the sample with full TES information (n=106) and the sample with imputed values for missing teachers (n=125). The set of estimates for the sample with full information reduces the number of classrooms for which we have a TES-score from 99 to 88.

Pupils can have multiple teachers during a school year. This is often the result of part-time appointments and a large share of female teachers, which is very common in primary education in the Netherlands. In case a class of children has more than one teacher, we weigh the scores on the TES by the number of teaching days. For instance, if teacher *X* teaches three days a week in class *Z*, and teacher *Y* the other two days, the TES-score for class *Z* is equal to (3/5)*TES teacher *X*+(2/5)*TES teacher *Y*.

In Figure 1 we show the distribution of the standardized TES-scores for our main sample (n=99). Standardization has been done by subtracting the mean (52.41) from the original score and dividing it by the standard deviation (13.28) such that the standardized TES-score has mean 0 and standard deviation 1. The distribution is skewed to the left. The 25th percentile of the standardized distribution is equal to -0.63, the median is equal to -0.07 and the 75th percentile is equal to 0.82. The minimum is equal to -2.74 and the maximum is 1.51.

We add covariates to the model. Our most important covariate is the previous test score derived from the end of school year 2010/2011. This previous test score is included as a control for ability differences between pupils. We include a second degree polynomial of this variable in our models. The previous test scores contain missing values because some pupils only entered the particular school in 2011/2012. We put missing test scores to zero. This is equal to the average of a particular school year and grade because of standardization of the data. We also include an indicator in the regression model when the previous test score is missing. Besides controlling for previous test scores, we control for other differences between pupils by including a second degree polynomial in age and dummies for gender, nationality, living in a one parent family, retention, and educational level of the parents. In our most comprehensive specification, we also control for observable classroom and teacher differences by including teacher experience,[8] class size, the average of the previous test scores, average age, fraction of girls, fraction of pupils with Dutch nationality, fraction of pupils living in a one parent family, fraction of pupils that retained, fraction of pupils with low-educated parents,[9] a dummy for classrooms that span multiple grades and a dummy for classrooms that have multiple teachers. We also include school and grade-fixed effects.

Table 1 provides descriptive statistics of our variables. Panel A presents means and standard deviations of pupil characteristics based on the sample for which grammar test scores are available. The fraction of pupils from low-educated parents equals 38 percent, which exceeds the average of 26 percent in this city. Almost half of the pupils live in a one parent family.

[8] Teacher experience has been weighted in the same way as the TES-score for a classroom. We define this as the teacher experience a classroom of children is confronted with.
[9] That is, parents who only finished the lowest level of secondary school or less.

Panel B shows descriptive statistics of classroom characteristics for the 99 classrooms for which TES-scores are available. The average of the unstandardized TES-score equals 52.41, with a standard deviation of 13.28. The average class size is 24. Panel C shows teacher characteristics. The vast majority of teachers (88 percent) is female, a usual situation in Dutch primary education. Average work experience amounts to 19 years, with 13 percent of all teachers having five years or less of work experience. Only 2 percent of the teachers in our sample have obtained a university degree. The remaining 98 percent has a degree in higher vocational education, which is the standard requirement for becoming a teacher in primary education in the Netherlands. In table A.2 in the Appendix we present all pairwise correlations between our class room variables. Interesting is that the (classroom) TES-score is not significantly correlated with (classroom) teacher experience: the correlation is -0.13 and the corresponding p-value 0.19.

3. Empirical strategy

The main goal of the empirical analysis is to estimate the relationship between TES-scores and pupil performance.[10] We employ a similar value-added type of model as used by Rockoff & Sperroni (2010) and Kane et al. (2011). These types of models account for the fact that teachers and pupils are not randomly assigned to classes within schools. We estimate a model in which the standardized test scores are related to standardized TES-scores in the following way:

$$\text{(M.1)} \quad Y_{ic} = \beta_0 + \beta_1 TES_c + \beta_2 Y_{t-1,ic} + \beta_3 Y_{t-1,ic}^2 + \beta_4' \boldsymbol{X}_{ic} + \beta_5' \boldsymbol{C}_c + \varphi_s + \theta_g + \varepsilon_{ic},$$

where Y_{ic} is the standardized test score of pupil *i* in school *s* in grade *g* in classroom *c*, and the previous test score is represented by $Y_{t-1,ic}$. $\boldsymbol{X}_{ic}$ is a vector of all other pupil characteristics and $\boldsymbol{C}_c$ is a vector consisting of all other classroom characteristics. To ease notation, we leave out indices for schools *s* and grades *g* for pupil and classroom variables. The term φ_s represents school-fixed effects (6 dummies, because we have 7 schools) and θ_g represents grade fixed effects (7 dummies, because we have 8 grades). Note that, by including the school and grade-fixed effects, we use variation between classroom variables within grades within schools. The parameter of interest is β_1, which represents the change in the test score of the pupil if the TES-score of the teacher changes. The estimated coefficients can be interpreted in terms of standard deviations.

[10] The goal of this paper is to investigate whether the total TES-score identifies teacher quality differences. We also investigated to what extent we could differentiate between (subsets of) competences by including them in the regressions simultaneously. However, disentangling factors of competence is difficult due to problems of multicollinearity as we work with only 99 classrooms and (highly) correlated competences. A principle component analysis reveals that the first component explains about 25% of the variance, and that the first 36 components of the 75 items explain about 90%. However, we could not give clear interpretations of the constructed (principle) components, which kept us away from using them in our analysis.

The 'value-added' type of model as presented in (M.1) should account for nonrandom assignment of teachers and pupils to classes. Rothstein (2010) criticizes these models because unobserved pupil characteristics make classrooms more easy or difficult to teach. This could play a role in assigning teachers to classes, which could yield biased estimates. Although these unobservable characteristics could confound the estimates of β_1, Kane & Staiger (2008) and Nye et al. (2004) show that experimental estimates - in situations where teachers have been randomly assigned to classrooms - are consistent with value-added estimates that result from the non-experimental value-added models. Also, Chetty et al. (2011) find no bias in value-added estimates using previously unobserved parent characteristics and a quasi-experimental research design based on changes in teaching staff. In the next section we investigate to what extent our results are affected by nonrandom assignment of teachers to classes based on a large number of observable pupil and classroom characteristics.

Another concern is that test scores are derived from the same school year as the TES-scores. This contemporaneous measurement could potentially confound the estimates of β_1 if there are unobserved class characteristics that independently affect both an evaluator's measurement and pupil achievement (Kane et al., 2011). For instance, if an evaluator encounters a teacher in a classroom with a high level of social cohesion, he may evaluate this teacher differently than he would have done if he encountered the same teacher in a classroom with a low level of social cohesion. At the same time, higher social cohesion may result in positive peer effects that raise pupil achievement, causing the estimates of β_1 to be biased. Kane et al. (2011) propose to use pupil achievement data from the previous or next year compared to the year the evaluations are carried out. Unfortunately, pupil achievement data from other years are not available to us. However, Tyler et al. (2009) show that results are in the same order of magnitude when same year pupil achievement data are used instead of previous or next year data. Moreover, we expect that this potential problem will be mitigated by the fact that the evaluators were educational professionals that were trained specifically for the evaluation job, and that we control for a rather large number of observable class characteristics that might be correlated with both the teacher TES score and pupil test scores, such as the average previous test score and the fraction of pupils from a one parent family. Hence, we believe that it is unlikely that our findings are spurious.

4. Estimation results

Tables 2, 3 and 4 present the main estimates of the relationship between the standardized TES-score and pupils' math, grammar and reading achievement, respectively. Each table has 5 columns, which include different sets of covariates. In column (1) we present the association between the TES-score and the test scores without any controls; in column (2) we add school and grade fixed effects; in column (3) we include previous test scores and other pupil characteristics; in column (4) we add all other classroom information and in column (5) we also include teacher experience. By including the school and grade-fixed effects we obtain an indication of the extent to which our results are affected by nonrandom sorting of teachers

across schools. By adding the previous test scores and the classroom variables we obtain an indication of the extent to which our results are affected by (possible) nonrandom sorting of pupils to classes *within* schools.

4.1. Basic estimates

The estimated coefficient for the TES-score change when including the school and grade-fixed effects, but the direction of the change is different for math than for reading or grammar (i.e., from 0.120 to 0.087 for math and from 0.078 to 0.123 for grammar). The direction of the change differs also between test domains when adding previous test scores and class variables: while for math the estimated coefficients increases from 0.087 in column (2) to 0.173 in column (3) and decreases to 0.143 in column (4), the coefficient for grammar drops from 0.123 to 0.103 and increases from 0.103 to 0.152, respectively. These changes in the coefficients suggest that pupils and teachers are assigned to classes in a nonrandom manner. However, given the difference in direction of the change between test domains, there is no indication that good teachers are generally assigned to either particular good or weak classes.

In any case, regardless of the type and direction of sorting, it does not seem to affect the statistical significance of the estimated coefficients for math and grammar. They are statistically significant in all columns of Tables 2 and 3. This suggests that these associations are robust to the inclusion of a range of covariates. When we control for teacher experience, the estimates slightly rise because of the negative but not significant correlation between teacher experience and the TES score (correlation=-0.13, p-value=0.19)

Based on the results in column (5) with all relevant controls, which is our preferred specification, we find that a higher TES-score is associated with higher pupil achievement scores for all three test domains and that this association is statistically significant. For math we find that, on average, a pupil gains 0.15 standard deviation if he is assigned a teacher that has a one standard deviation higher score on the rubric. For grammar the estimated gain is about 0.18 standard deviations. The estimated coefficient for reading is 0.11 standard deviations. This coefficient points at a somewhat weaker association with reading scores, however still significant at a 5-percent significance level in the model with all controls.

4.2. Nonlinear effects

The TES-score has been treated linearly in the analyses presented so far. To investigate the possibility of a nonlinear relationship we split up the TES-score in quartile dummies. Table 5 presents the results of regressions in which dummies for the quartiles of the TES-score have been included instead of the linear TES-score. The presented specification includes all covariates.

The estimates suggest that replacing a teacher from the lowest quartile of the TES-score distribution by a teacher from the upper quartile yields test score gains of 0.37 standard

deviations in math, 0.44 in grammar, and 0.24 in reading. These gains are relatively large.[11] They suggest that a pupil who starts the year at the 50th percentile and is assigned to a teacher from the highest quartile is likely to gain on average 15 percentile points more in math, 19 in grammar, and 10 in reading compared to a pupil assigned to a teacher from the lowest quartile.

The TES-rubric seems to particularly differentiate between the weakest teachers and the rest. The relationship between the other quartiles is less clear. The point estimate increases from the second quartile to the top quartile, but differences between the second, third and top quartile are not statistically significant. This suggests that our results are especially relevant for the weakest performing teachers on the rubric. Holtzapple (2003) and Kane et al. (2011) find similar results for Cincinatti's TES system.

4.3. The impact of missing end of year evaluation scores

Next, we conduct an analysis on the set of teachers for which we have both start and end year evaluation scores. This means that we exclude from the sample those teachers for which no end score on the rubric is available. This reduces our sample of teachers from 125 to 106 and our sample of classrooms from 99 to 88.

Table 6 presents a set of estimates in line with the specification reported in column (5) of Tables 2-4. The estimated coefficients do not change compared to the ones obtained in the specification in which we have used imputed values. We conclude that our estimated coefficients are unlikely to be influenced by the imputation of teachers' average evaluation score by their start of year evaluation score.

5. Conclusion

This research reports the results of a program aimed at improving teacher performance in primary education in a large city in the Netherlands. We obtain a set of estimates suggesting that teachers with higher evaluation scores, on a detailed classroom observation instrument, produce greater average gains in pupil achievement, measured by pupil test scores. Estimates of these gains range from 0.11 (reading) to 0.19 (grammar) standard deviations if a pupil is assigned to a teacher with a one standard deviation higher evaluation score on the rubric. This finding is consistent with prior work in the United States.[12]

In addition, we find that the rubric is particularly successful in distinguishing weak teachers from other teachers, but less so in differentiating between an average and an excellent teacher. This observation is also consistent with US findings. Estimated gains from being

[11] Comparable estimates in Kane et al. (2011) for Cincinatti's Teacher Evaluation System are 0.09 for Math and 0.13 for reading.
[12] Comparable estimates for Cincinatti's TES are 0.09 in math and 0.08 in reading (Kane et al., 2011). Rockoff & Speroni (2010) report a coefficient of 0.05 higher math achievement of a one standard deviation higher rating by teacher mentors.

assigned to a teacher in the highest quartile instead of a teacher in the lowest quartile in the evaluation rubric are between 0.24 to 0.44 standard deviations.

These results suggest that evaluations made by trained experts on a detailed rubric seem to have potential to address the problem of weak teacher quality. One of the advantages of using these rubrics with detailed standards for teacher practices over more subjective ratings is that the score on the rubric provides signals to teachers and principals as to in what particular competences or classroom practices improvements can be made. This information could be effectively used in personal development plans to improve teacher quality. Promising in this respect is that Taylor & Tyler (2011) show that repeated evaluations and feedback to (mid-career) teachers by trained experts based on a detailed rubric raise pupil achievement, particularly in the years after the evaluation and feedback have been carried out.

While our results show statistically significant correlations between the overall set of classroom practices and pupil achievement, we cannot rule out the possibility that the true causal relationship is different. The main reason to be cautious is that we have not been able to rule out all biases due to nonrandom matching of teachers to classes. Further experiments with coaching on particular classroom practices are likely to shed more light on which (types of) practices are most crucial for assessing and improving teacher quality and pupil performance.

A last point of attention for any evaluation system is the reliability of the evaluations when carried out by different evaluators. Rockoff & Speroni (2010) find variation in the leniency between evaluators, particularly in the case of evaluations by mentors. This problem is likely to be reduced when using independent evaluators, who probably have fewer incentives to be lenient. Kane & Staiger (2012) conclude that it seems possible to constrain tendencies to score too lenient or too harsh when training of evaluators is taken seriously. In sum, evaluating teachers by classroom observations, using a detailed rubric conducted by trained evaluators, is likely to be a promising avenue for education policies that aim to improve teacher performance in primary education.

Literature

Aaronson, D., L. Barrow, and W. Sander, 2007, Teachers and Student Achievement in the Chicago Public High Schools, *Journal of Labor Economics*, vol. 25, no. 1, pp. 95–135.

Chetty, R., N. Friedman, and J. Rockoff, 2011, The long-term impact of teachers: teacher value-added and student outcomes in adulthood, *NBER Working Paper*, no. 17699.

Clotfelter, C., H. Ladd, and J. Vigdor, 2006, Teacher–Student Matching and the Assessment of Teacher Effectiveness, *NBER Working Paper*, no. 11936.

Hanushek, E., and S. Rivkin, 2010, Using Value-Added Measures of Teacher Quality, *American Economic Review*, vol. 100, no. 2, pp. 267–71.

Holtzapple, E., 2003, Criterion-related validity evidence for a standards-based teacher evaluation system, *Journal of Personnel Evaluation in Education*, vol. 17, no. 3, pp. 207-219.

Jacob, B., 2007, The Challenges of Staffing Urban Schools with Effective Teachers, *The future of Children*, vol. 17, no. 1, pp. 129–54.

Jacob, B., and L. Lefgren, 2008, Principals as agents: Subjective performance measurement in education, *Journal of Labor Economics*, vol. 26, no. 1, pp. 101-136.

Kane, T., and D. Staiger. 2008, Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, *National Bureau of Economic Research Working Paper*, no. 14601.

Kane, T., E. Taylor, J. Tyler, and A. Wooten, 2011, Identifying effective classroom practices using student achievement data, *Journal of Human Resources*, vol 46, no. 3, pp. 587-613.

Kane, T., and D. Staiger, 2012, Gathering feedback for teaching: combining high-quality observations with students surveys and achievement gains, *Measures of Effective Teaching Research Paper*.

Nye, B, S. Konstantopoulos and L. Hedges, 2004, *Educational Evaluation and Policy Analysis*, vol. 26, no. 3, pp. 237-257

Rivkin, S., E. Hanushek, and J. Kain. 2005, Teachers, Schools and Academic Achievement, *Econometrica*, vol. 73, no. 2, pp. 417–58.

Rockoff, J., 2004, The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data, *American Economic Review*, vol. 94, no. 2, pp. 247-252.

Rockoff, J., and C. Speroni, 2010, Subjective and objective evaluations of teacher effectiveness, *American Economic Review, Papers & Proceedings*, vol. 100, pp. 261-266.

Rothstein, J., 2010, Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement, *Quarterly Journal of Economics*, vol. 25, no. 1, pp. 175-214.

Staiger, D., and J. Rockoff, 2010, Searching for effective teachers with imperfect information, *Journal of Economic Perspectives*, vol. 23, no. 3, pp 97-118.

Taylor, E., and J. Tyler, 2011, The effect of evaluation on performance: evidence from longitudinal student achievement data of mid-career teachers, *NBER Working Paper*, no. 16877.

Tyler, J., E. Taylor, T. Kane and A. Wooten, 2009, Using student performance data to identify effective classroom practices, Working Paper.

Tyler, J., E. Taylor, T. Kane, and A. Wooten, 2010, Using student performance data to identify effective classroom practices, *American Economic Review Papers and Proceedings*, vol. 100, pp. 256-260.

Figure & Tables

Figure 1: The distribution of the standardized class room teacher evaluation score (N=99, i.e. number of classrooms)

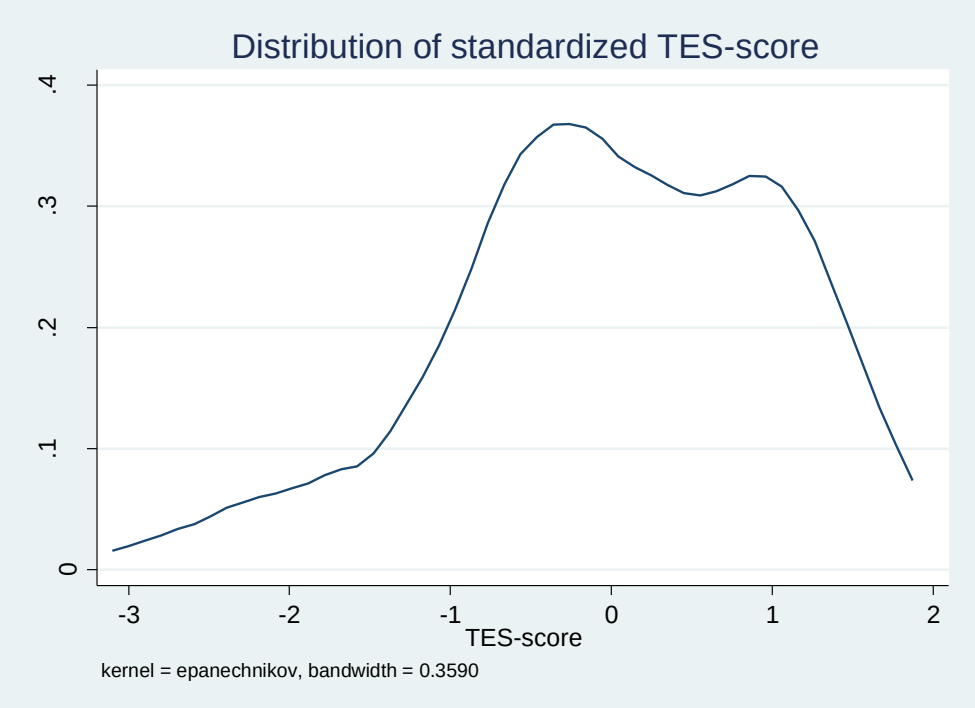

Table 1: Descriptive Statistics, restricted and unrestricted sample

	Unrestricted sample**		Restricted sample***	
Panel A: Pupil characteristics (based on grammar sample)	**mean**	**sd**	**mean**	**sd**
Girl	0.50	0.50	0.50	0.50
Age	8.05	2.40	8.69	2.35
Dummy=1 if low educated parents*	0.38	0.49	0.38	0.49
Dummy=1 if from one parent family	0.49	0.50	0.51	0.50
Dummy=1 if Dutch nationality	0.90	0.30	0.90	0.30
Dummy=1 if retained	0.07	0.25	0.06	0.23
Number of observations	2110		1859	
Panel B: Classroom characteristics	**mean**	**sd**	**mean**	**sd**
Classroom TES-score (unstandardized)	52.41	13.28	54.23	12.49
Classroom teacher experience	19.19	9.60	18.99	9.75
Class size	24.17	4.15	23.83	4.01
Classroom spans multiple grades (%)	20.20	40.35	18.18	38.79
Classroom has multiple teachers (%)	33.33	47.38	36.36	47.38
Fraction of girls (%)	50.92	7.74	50.99	7.87
Average age	8.01	2.41	8.08	2.38
Fraction of pupils with low educated parents (%)	38.69	12.64	38.82	12.23
Fraction of pupils from one parent family (%)	50.56	14.79	51.32	14.51
Fraction of pupils with Dutch nationality (%)	88.02	9.197	88.42	8.23
Fraction of pupils that retained (%)	6.20	12.10	5.56	11.91
Number of classrooms	99		88	
Panel C: Teacher characteristics	**mean**	**sd**	**mean**	**sd**
Female (%)	88.24	32.37	89.62	30.64
Higher vocational education as highest level of educational attainment (%)	98.02	14.00	97.59	15.43
Experience in education (years)	19.45	11.61	19.68	11.53
Five years or less experience in education (%)	12.75	33.51	11.90	32.58
In higher pay scale (%)	7.14	25.88	8.54	28.11
Tenure (%)	92.16	27.02	95.24	21.42
Size of contract (% of FTE)	87.19	18.16	88.03	17.69
Number of teachers	125		106	

* Finished lowest track of secondary school or less. ** The unrestricted sample is the sample of teachers and their classes for which the start-of-the-year teacher evaluation score is available. *** The restricted sample is the sample of teachers and their classes for which both start and end of year teacher evaluation score is available.

Table 2: Relationship between TES and math score

	Dependent variable: standardized math score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
TES-score	0.120***	0.087**	0.173***	0.143***	0.154***
	(0.040)	(0.040)	(0.050)	(0.051)	(0.051)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2084	2084	2084	2084	2084
Number of classrooms	99	99	99	99	99
R-squared	0.017	0.054	0.416	0.440	0.449

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Table 3: Relationship between TES and grammar score

	Dependent variable: standardized grammar score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
TES-score	0.078*	0.123***	0.103**	0.152***	0.178***
	(0.041)	(0.046)	(0.047)	(0.047)	(0.046)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2110	2110	2110	2110	2110
Number of classrooms	99	99	99	99	99
R-squared	0.009	0.038	0.333	0.365	0.375

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Table 4: Relationship between TES and reading score

	Dependent variable: standardized reading score				
Independent variable:	(1)	(2)	(3)	(4)	(5)
TES-score	0.039	0.076	0.034	0.083	0.107**
	(0.037)	(0.046)	(0.045)	(0.050)	(0.050)
School and grade fixed effects	no	yes	yes	yes	yes
Previous test scores and other pupil characteristics	no	no	yes	yes	yes
Classroom variables	no	no	no	yes	yes
Teacher experience	no	no	no	no	yes
Observations	2135	2135	2135	2135	2135
Number of classrooms	99	99	99	99	99
R-squared	0.002	0.026	0.366	0.387	0.398

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Table 5: Relationship between quartiles of teacher TES-score and pupil math, grammar and reading scores

	Dependent variable:		
	math	grammar	reading
Independent variable:	(1)	(2)	(3)
Indicator for TES-score between 25th and 50th percentile	0.326***	0.399***	0.149
	(0.0967)	(0.0960)	(0.109)
Indicator for TES-score between 50th and 75th percentile	0.290***	0.337***	0.248**
	(0.107)	(0.104)	(0.120)
Indicator for TES-score between 75th and 100th percentile	0.371***	0.440***	0.236*
	(0.115)	(0.112)	(0.120)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	2084	2110	2135
Number of classrooms	99	99	99
R-squared	0.453	0.381	0.399

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Table 6: Relationship between teacher TES-score and math, grammar and reading, restricted sample

	Dependent variable:		
	math	grammar	reading
Independent variable:	(1)	(2)	(3)
TES-score	0.145***	0.153***	0.0887
	(0.0511)	(0.0483)	(0.0576)
School and grade fixed effects	yes	yes	yes
Previous test scores and other pupil characteristics	yes	yes	yes
Classroom variables	yes	yes	yes
Teacher experience	yes	yes	yes
Observations	1833	1859	1863
Number of classrooms	88	88	88
R-squared	0.447	0.370	0.391

Standard errors clustered on classroom between brackets, *** p<0.01, ** p<0.05, * p<0.1

Appendix

Table A.1: The teacher evaluation rubric

Indicator	Type	Subitems
Clearly sets high expectations	p	4
Instruction takes [adequate] account of [relevant] differences between pupils	p	4
Assimilation of lesson material takes account of differences between pupils	p	4
Provides extra instruction and time to learn for weaker pupils	p	3
Makes clear how the lesson fits in with earlier lessons	d	4
Clearly states the lesson goals at the beginning of the lesson	d	3
Provides insight into the organization of the lesson	d	3
Clearly explains the lesson material and assignments	d	4
Provides feedback to pupils	d	6
Checks that lesson goals have been reached	d	5
Stimulates reflection via interactive instruction and work methods	d	2
Encourages pupils to think out loud	d	2
Teaches pupils strategies for thinking and learning	d	6
Encourages pupils to reflect on differing solution strategies	d	5
Encourages the use of control activities (checks)	d	3
Stimulates application of the lesson material	d	3
Spends the planned time on the [actual] lesson goals	o	5
Ensures the lesson follows an adequate planning	o	9
	Total	75

p = pedagogical competence; d = didactical competence; o = organizational competence

Table A.2: Matrix of pairwise correlations between class room variables (n=99). P-values in italics

Nr.	Description:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Class room TES-score	1,00																
2	Class room teacher experience	-0,13	1,00															
		0,19																
3	Class size	-0,13	-0,03	1,00														
		0,19	*0,74*															
4	Classroom spans multiple grades (%)	-0,23	0,01	0,17	1,00													
		0,02	*0,91*	*0,08*														
5	Classroom has multiple teachers (%)	0,29	-0,15	-0,16	0,02	1,00												
		0,00	*0,13*	*0,12*	*0,86*													
6	Fraction of girls (%)	-0,09	0,00	-0,14	-0,06	-0,14	1,00											
		0,36	*0,99*	*0,18*	*0,57*	*0,17*												
7	Average age	0,22	-0,07	-0,14	-0,24	0,19	-0,09	1,00										
		0,03	*0,50*	*0,17*	*0,02*	*0,06*	*0,40*											
8	Fraction of pupils with low educated parents (%)	0,11	-0,02	0,03	0,02	-0,03	-0,09	0,41	1,00									
		0,27	*0,81*	*0,76*	*0,84*	*0,79*	*0,38*	*0,00*										
9	Fraction of pupils from one parent family (%)	0,16	0,00	-0,16	-0,05	0,14	-0,09	0,27	0,15	1,00								
		0,11	*0,97*	*0,11*	*0,60*	*0,16*	*0,37*	*0,01*	*0,15*									
10	Fraction of pupils with Dutch nationality (%)	-0,03	0,23	-0,09	-0,39	-0,15	-0,03	0,05	-0,09	0,11	1,00							
		0,80	*0,02*	*0,38*	*0,00*	*0,14*	*0,74*	*0,62*	*0,36*	*0,28*								
11	Fraction of pupils that retained (%)*	-0,15	0,01	0,22	0,33	-0,08	0,08	-0,55	-0,17	-0,16	-0,13	1,00						
		0,15	*0,94*	*0,03*	*0,00*	*0,41*	*0,45*	*0,00*	*0,10*	*0,11*	*0,21*							
12	Average test score math	0,27	0,01	0,05	-0,12	-0,08	0,33	-0,08	0,03	-0,09	-0,05	0,14	1,00					
		0,01	*0,96*	*0,61*	*0,25*	*0,44*	*0,00*	*0,45*	*0,77*	*0,35*	*0,60*	*0,17*						
13	Average test score grammar	0,17	0,09	0,10	-0,02	-0,21	0,29	-0,04	0,04	-0,05	0,03	0,18	0,73	1,00				
		0,10	*0,35*	*0,30*	*0,85*	*0,04*	*0,00*	*0,71*	*0,71*	*0,64*	*0,74*	*0,08*	*0,00*					
14	Average test score reading	0,07	0,02	-0,11	0,09	-0,01	0,17	0,01	0,03	0,00	-0,03	0,20	0,43	0,62	1,00			
		0,50	*0,86*	*0,29*	*0,36*	*0,89*	*0,10*	*0,90*	*0,74*	*0,98*	*0,74*	*0,05*	*0,00*	*0,00*				
15	Average previous test score math	0,00	-0,15	0,06	-0,28	-0,12	0,05	0,06	-0,12	-0,10	0,02	-0,18	0,32	0,10	-0,03	1,00		
		0,98	*0,15*	*0,56*	*0,00*	*0,23*	*0,61*	*0,55*	*0,22*	*0,34*	*0,84*	*0,07*	*0,00*	*0,32*	*0,77*			
16	Average previous test score grammar	-0,03	-0,06	0,25	-0,12	-0,29	0,14	0,05	0,15	-0,15	-0,11	-0,08	0,30	0,33	0,10	0,39	1,00	
		0,75	*0,55*	*0,01*	*0,24*	*0,00*	*0,16*	*0,64*	*0,14*	*0,14*	*0,27*	*0,44*	*0,00*	*0,00*	*0,33*	*0,00*		
17	Average previous test score reading	-0,03	-0,13	0,22	0,01	-0,16	0,13	0,09	0,21	-0,01	-0,24	-0,06	0,21	0,21	0,17	0,18	0,78	1,00
		0,77	*0,19*	*0,03*	*0,90*	*0,11*	*0,20*	*0,40*	*0,03*	*0,90*	*0,02*	*0,53*	*0,04*	*0,03*	*0,10*	*0,08*	*0,00*	

* Most pupils retained in grade 1: they enrolled in the school year when they turned four and stayed an extra year in kindergarten

Publisher:

CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510 | 2508 GM The Hague
T (070) 3383 380

January 2013 | ISBN 978-90-5833-581-4