Performance-Related Pay for Teachers in the Netherlands:

from basic theory to teachers support

ERASMUS UNIVERISTY ROTTERDAM

Erasmus School of Economics

Department of Economics

Supervisor: M.W.J.M. Buurman

Name: Julian Opdam

Exam number: 302675

E-mail address: julian_o@live.nl

Abstract

Recently there seems to be an increased interest in performance-related pay for teachers. New programs have been initiated in the United States, England and Portugal. In the Netherlands interest is rising too, as the newly formed cabinet Rutte I has decided to structurally invest in performance-related pay of teachers in the near future. This thesis will try to answer if the introduction of performance-related pay for teachers is indeed a good idea. The basis theory of performance-related pay is discussed as well as the current empirical evidence. The main new contribution of this thesis will be the analysis of teachers support for performance-related pay. A survey of the Dutch Socialist Party (SP) is used to investigate this. With the help of regressions it is investigated if school size, urbanization, school type, gender or age has any effect on teachers support. The main finding is that younger teachers (20-30 years old) are significantly more positive toward performance-related pay than older teachers. This might be a welcoming message for policymakers as in the coming years many old teachers will leave the profession and will be replaced by younger ones. Overall the introduction of a good performance-related pay program for teachers will be challenging but not impossible.

Acknowledgements

The completion of this thesis will mark the ending of my studies for the master program Economics of Markets, Organizations and Policy at the Erasmus University Rotterdam. I would not have been successful without the help of certain people.

First, I want to thank my thesis supervisor Margaretha Buurman. You often provided helpful comments and was always quick to reply to my e-mails. When discussing my thesis in person you always took your time, and never rushed. I highly appreciated that. Second, I would like to thank the Socialistic Party (SP) for giving me access to the results of their survey. Specifically, I would like to thank Manja Smits who was my contact person in the SP.

Of course this acknowledgement cannot end before thanking my parents. You always supported me and showed interest in my studies. I most likely would not have made it this far without your support.

Julian

Table of Content

Chapter 1: Introduction

"Good teachers are costly, but bad teachers cost more."

- Bob Talbert

Performance-related pay (PRP), talking about it while the financial crisis of 2007 is still raging on may seem strange. It is no secret that the billions[1] spend on bonuses played a big part in the realization of the financial crisis. Above all, it provided a clear warning that care is needed when using performance-related (monetary) incentives. However, dismissing performance-related pay as faulty and unusable is naïve. It is well established in the economic literature that people respond to incentives, both theoretical as empirical, see for example Gibbons (1998), Predergast (1999), and Jenkins et al. (1998).

Recently, policy makers seem to have an increased interest in incentive programs for teachers. In the United States new programs have been initiated under the No Child Left Behind act, in Portugal and the United Kingdom educational reforms have introduced performance-related pay parts into the teachers pay scale. Even here, in the Netherlands, interest is rising. The newly formed cabinet Rutte I declared in the coalition agreement to structurally invest €250 million in performance-related pay, starting in 2015.

This recent interest in performance-related pay programs may be explained by (new) findings that emphasis the importance of teachers (see chapter 3). With decreasing scholastic performance of students, shortages of teachers and the current importance of the knowledge economy in mind, policymakers try to find new ways to improve the educational system.

Proponents of performance-related pay argue that it can help solve these problems as it may attract, retain and motivate high quality teachers. Moreover they argue that failure of past incentive plans (in the United states) was due to teacher and union opposition (Ballou, 2001). Opponents respond by claiming that the failure was due difficulties in identifying the most effective teachers. They therefore conclude that the type and nature of work that is done by teachers makes the use of performance-related pay for teachers impossible (Murnane & Cohen, 1986). However, Ballou (2001) finds that private schools widely make use of performance-related pay. Countering the idea that the teaching profession is the problem. He concludes that rather specific circumstance in public education are to blame. In a recent study Woessmann (2010) finds that students from countries that use performance-related teacher

[1] The New York State comptroller found that cash bonuses paid by Wall street firms in 2008 accrued to $18.4 billion dollar. This was a sharp decline relative to 2007 when $32.9 billion of bonuses were paid. http://osc.state.ny.us/press/releases/jan08/bonus.pdf (accessed 6 October 2010)

pay score significantly higher than students from countries without performance-related teacher pay.

With the intention of the Dutch government in place to structurally invest in performance-related pay for teachers, now is the time to determine if this is indeed a good investment to make. Moreover, as the Ministry of Education, Culture and Science does not has any specific plans yet for a performance-related pay program[2]. The main question this thesis therefore tries to answer is:

Should performance-related pay for teachers be introduced in the Netherlands?

The main (new) contribution of this thesis will be to investigate the support of Dutch teachers for performance-related pay. This is done by analyzing a survey that the Dutch Socialist Party (SP) held under teachers in 2009. The analysis contains 967 teachers from secondary education and 1426 teachers from primary education. Investigating teachers support is interesting as it will highlight possible bottlenecks from the viewpoint of teachers. For example, the results indicate that teachers have little faith in a transparent assessment from their principal. Including such an assessment in a possible performance-related pay program may therefore be unwanted as it may demotivate a large proportion of teachers. By my knowledge only Marsden (2000) has done a similar analysis. He investigated teachers support before the introduction of a performance-related pay reform in the United Kingdom (see also section 7.2.3). Furthermore, this thesis will, by using a regression analysis, also investigate if different teachers differ in their support for performance-related pay. For example, are male teachers more supportive towards performance-related pay than female teachers? This will allow us to be even more specific towards possible problem areas for an incentive program. This may help to indicate under which teachers possible pilot programs may be best started up or were to direct extra attention when designing and implementing a possible future performance-related pay program. To my knowledge no other paper as done such an analysis. Although, Marsden (2000) does on some questions compares the answers between younger and older teachers, he does not employ a regression analysis.

[2] See for example an interview with Minister Van Bijsterveldt: http://nos.nl/video/190827-van-bijsterveldt-wordt-minister.html

The main finding of our analysis is that younger teachers, specifically in the age category of <20-30 and to a lesser extend in the age category of 30-40, are significant more supportive of performance-related pay than older teachers. School size, urbanization, and gender do not seem to have any impact on teachers support. This may be a welcoming message to policymakers as in the near future many old teachers will leave the profession and will be replaced by younger ones.

The remainder of this thesis is structured as followed. Chapter 2 will start by describing the current and future projected state of the Dutch labor market for teachers. Chapter 3 will discuss the importance of teachers. How important are they for students test results? Which characteristics do good teachers have? This will help us explain the shortcomings of the current pay system. In chapter 4, I will explain the basic theory of performance-related pay. What are the advantages over current input based pay? Which problems may arise when introducing performance-related pay? In chapter 5, I will specifically look at problems that may arise from the interaction with psychological factors. Intrinsic motivation will be discussed in-depth but also other factors like reciprocity, fairness and social norms will be considered. Chapter 6 will discuss how teachers performance may be measured. What are the pros and cons of objective and subjective measures? In chapter 7 the current existing empirical evidence is discussed. How successful have previous programs been? What can we learn from them? Chapter 8 analysis a data set from the Socialistic Party (SP) that consist of a survey held under Dutch teachers. Do Dutch teachers support the idea of performance-related pay? Do different kind of teachers differ in their opinion? Chapter 9 will summarize and make some policy recommendations.

Chapter 2: Labor Market for Teachers

"Our progress as a nation can be no swifter than our progress in education. The human mind is our fundamental resource."

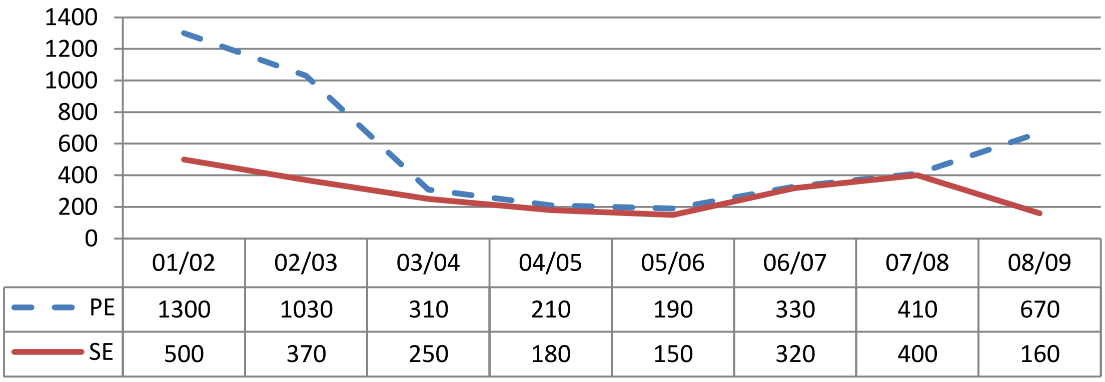
- John F. Kennedy

The Dutch labor market for teachers has been a concern for some years now. There are two main problems identified: the shortage of teachers, now and in the future, and the quality of teachers (de Commissie Leraren, 2007). The main concern is that if these problems are not solved the quality of the Dutch educational system will decrease. Which, in turn, will hurt the Dutch knowledge economy. In this section I will describe the magnitude of the current and future shortage of teacher and shortly address the quality problem. The development of the age distribution will also be described as well as the current salary structure of Dutch teachers.

2.1. Current shortage

The most important problem that the Dutch labor market for teachers faces is the shortage of teachers. The shortage of teachers in primary education is around 670 fulltime jobs in the school year of 2008/09, which is 0,7 percent of the total demand for teachers. This percentage is called the 'vacancy intensity'. For secondary education there is a shortage of around 160 fulltime jobs, which is around 0,2 percent of the total demand for teachers. See also figure 2.1 below. To put this vacancy intensity in perspective, a percentage above 1 percent will result in classes sometimes being send home because there is no teacher available (de Commissie Leraren, 2007).

Figure 2.1: Average outstanding vacancies (in FTE) in corresponding school year



	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09
PE	1300	1030	310	210	190	330	410	670
SE	500	370	250	180	150	320	400	160

source: WIO 2010

2.2. Future shortage

Every few years an estimation is done for the labor market for teachers. Previous estimates painted a rather grim future. An estimation done in 2006 predicted that in 2015 vacancy intensities in primary education would rise to 3 percent. For secondary education vacancy intensities would even rise further to a minimum of 9 and a maximum of 18 percent in 2015 (Ecorys et al. 2006).

The newest estimate from 2010 predicts a more brighter future (OCW, 2010). The shortage for primary teachers will almost disappear in the coming years. From 2014 it will steadily rise again to a shortage of around 1000 fulltime jobs in 2020, this corresponds to vacancy intensity of approximately 1 percent. In secondary education a reverse pattern is estimated. In the coming years the shortage will rise 2600 fulltime jobs in 2016. It will then sharply drop to no shortage of teachers in 2020. This drop is mostly driven by a decrease in the amount of students, which lessens the demand for teachers.

The differences between the two predictions are mainly caused by assumptions that the new model makes regarding the working of public policy (i.e. actieplan *Leer*kracht van Nederland). Furthermore, the financial crisis, and it's future consequences, also dampens the shortage of teachers. More teachers stay in the profession as it is a relative safe job. To conclude it seems that the new prediction indicates that the shortage of teachers might be less of a problem than previous thought.

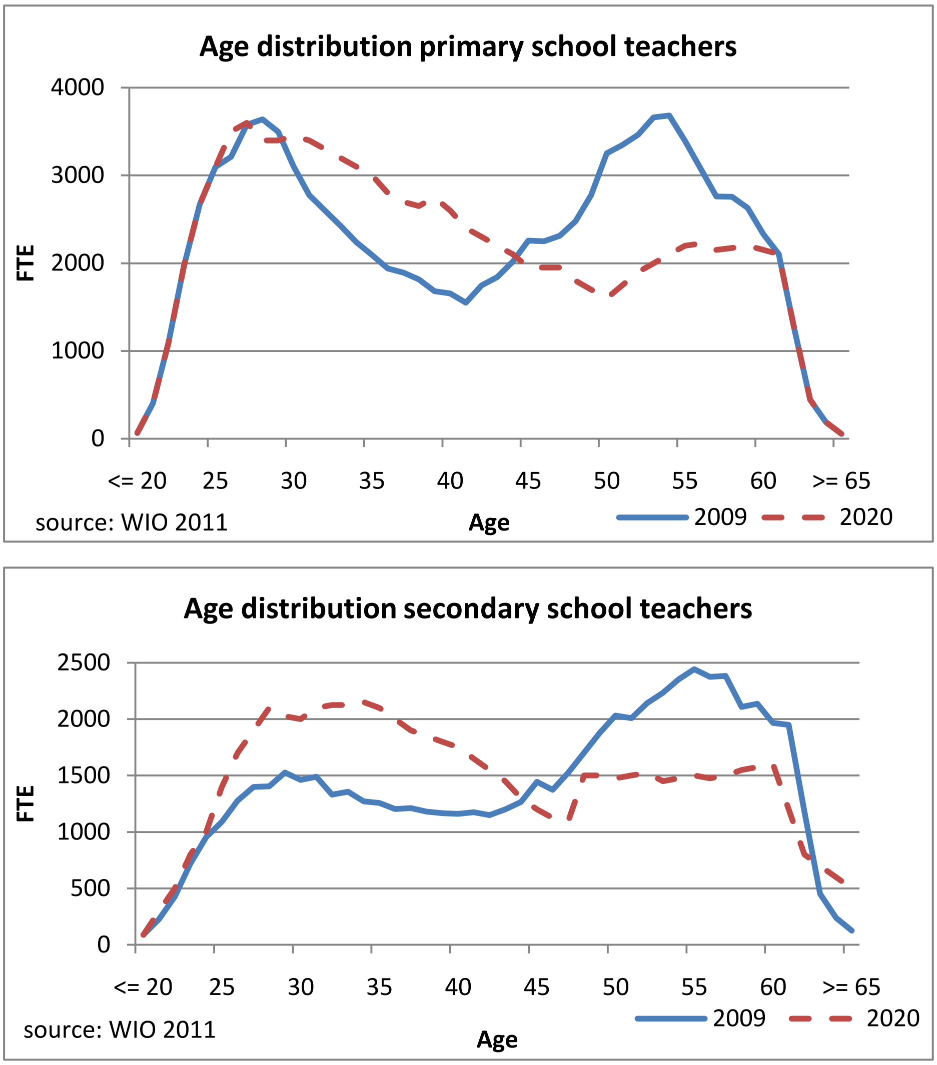
2.3. Age distribution

One of the reasons why the shortage in secondary education will rise the coming years is the highly aged teacher distribution in the Netherlands. In secondary education 46 percent of the teachers are 50 years or older, for primary education this is 35 percent. In comparison, the average percentage for the complete Dutch labor market is only 25 percent. Indeed, the education sector is the most aged sector in the Dutch economy (de Commissie Leraren, 2007). In the future this age distribution will change dramatically. Many old teachers will leave the profession, with younger teachers replacing them. Figure 2.2 gives a overview.

2.4. Quality

Another problem that the labor market for teacher faces is a decrease in the quality of teachers. The quality of teachers is measured by their certification status. As we will see in section 2 quality and certification status are not quite the same. Making the notion that the

Figure 2.2: Age distributions



source: WIO 2011

source: WIO 2011

quality of teachers decreases somewhat misleading. At present one in six classes (16.9 percent) in secondary education is taught by an uncertified teacher (OCW, 2010). For primary education no systematic statistics are available. Samples indicate that around 1 percent of the classes in primary education is taught by an uncertified teacher (www.bevoegd.nl).

2.5. Payment

Currently teachers in the Netherlands get paid according to a single salary structure. Teachers are differentiated into categories, in secondary education these categories are: LB, LC and LD. Teachers in the LB and LC category need at least a degree from an university of applied sciences; this is a Dutch institution of higher education which is one step below a regular university. When finished they obtain a certificate of second degree teaching competence.

Which basically means that they can teach all classes except the last two/three years of the highest forms of secondary education (HAVO and VWO). To enter the LC category teachers need to follow additional courses. The highest category is LD, were teachers need a degree from an university, when finished they obtain a first degree of teaching competence. They may teach all classes in secondary education.

In 2010 around 65 percent of the secondary teachers are in the LB category, 17 percent are in LC category and 18 percent are in the LD category. The government aims to have this changed to 33 percent in LB, 38 percent in LC, and 29 percent in LD by 2014 (OCW, 2010). Teacher beginning salaries are subsequently €2445 for LB, €2460 for LC, and €2470 for LD. This increases in 17 steps to the final salary of €3739, €4361, and €4962 (VO CAO, 2010). This will also change the coming years as the government has decided to decrease the amount of steps to 12. Resulting in a faster progression in teachers salary. Normally these steps are automatically awarded annually and increases salary around 2 to 7 percent, depending on category and step number.

For primary education the salary scales consist of LA, LB, LC. To get certification teachers have to complete the PABO (a degree from an university of applied sciences). Ninety-eight percent of the teachers falls in LA category. The LA scale starts at a salary of €2270, and ends after 17 steps at €3247 (PO CAO, 2009). The intention is to decrease the amount of teacher in the LA category and promote them to the LB category. The government aims to have a 58 percent in LA, 40 percent in LB, and 2 percent in LC by 2014 (OCW, 2010).

To put this wage in perspective. The average wage of all people active on the Dutch labor market is around €2850 per month. The starting salary of teaching is considered good with the final salary teachers reach considered average. Trends in beeld 2010 (OCW, 2010)[3] indicates that after 15 years of experience primary and lower secondary teachers lag behind in wage relative to their counterparts with the same level of education. Teachers in higher secondary education earn more than their average counterparts.

Internationally, teachers pay in the Netherlands is below the top. Being ranked 9th of the 28 OECD countries included in the study of Woessmann (2010)[4]. If teachers salary is

[3] this specific information is found on: http://www.trendsinbeeld.minocw.nl/grafieken/3_1_3_24.php

[4] Woesmann obtained data about teacher salary levels in 2004 from the OECD (2006). He uses a measure of statutory public school teacher salary after 15 years of experience with minimal training, averaged over primary and lower secondary education.

taken relative to GDP per capita the Netherlands falls two places and is ranked 11th. Overall we can say that teachers in the Netherlands are not badly paid but also not particularly great.

Teachers themselves often do not seem very satisfied with their salary. In the survey of the SP (see further chapter 9) teachers could rate their satisfaction with their salary on a 5 point scale. The average of both primary as secondary teachers was a 2, noted as insufficient. Also in other surveys teachers are negative about their salary. In a survey held by the Ministry of the Interior and Kingdom relations (Personeels- en Mobiliteitsonderzoek, POMO) only around 20 percent of the teachers thinks that their salary is on the right level (table 2.1.A). Interestingly, only around 7 percent thinks that their individual performance matters for their pay. Seven percent also thinks that there is not enough growth potential in their salary (table 2.1.B). Hinting at possible openings for performance related pay.

Table 2.1.A	In comparison with other companies, teacher salary is on the right level	My salary fits the level of my function
Completely disagree	20.5 %	21.1 %
Disagree	25.7 %	30.3 %
Neutral	22.3 %	22.7 %
Agree	15.3 %	18.4 %
Completely agree	5.0 %	3.9 %
Source: stamos.nl, POMO 2005		

Table 2.1.B	My individual performance matters for my pay	There is enough financial growth potential
Completely disagree	43.0 %	36.6 %
Disagree	25.2 %	29.9 %
Neutral	14.8 %	18.1 %
Agree	5.9 %	5 %
Completely agree	1.4 %	1.7 %
Source: stamos.nl, POMO 2005		

Chapter 3: Importance of Teachers

“Teaching is the only major occupation of man for which we have not yet developed tools that make an average person capable of competence and performance. In teaching we rely on the ‘naturals’, the ones who somehow know how to teach.”

- Peter Drucker

Trying to increasing the quality of teachers is of course only truly relevant if teachers matter for the result of their students. The Coleman report (Coleman et al. 1966) surprised the world by concluding that family background and peers were more important than schools and teachers. Today, most researchers agree that schools and teachers do matter, see for example Hanushek and Rivkin (2006). In this section I will try to discover how much teachers matter to scholastic achievements of students and which characteristics define good teachers. Along the way we will discover why the discussion about PRP may be relevant.

Wayne and Youngs (2003) provide a review of 21 studies about the link between teacher characteristics and student achievements. It has three interesting findings. One, it does not find clear evidences that degree level (bachelor or master degree) of a teacher matters, except for math teachers. Two, it finds evidence that certification status (certification versus no certification) matters for math teachers, although this conclusion is based on only one study. That same study did not find differences between teachers with standard certification versus temporary, provisional or emergency certification. Three, there seems to be positive relationship between teacher test scores, measured for example by their score on their licensure exam or on a standardized high school tests, and student test scores. Unless the studies control for college ratings. Wayne and Youngs (2003) also provide results of some positive relationship between teachers from better-rated undergraduate institutions and student test scores.

Hanushek and Rivkin (2006) provide another review of the literature in their handbook of the Economics of Education. They also conclude that having a master’s degree does not seem to have a significant impact on students test scores. Teachers experience does also not seem to have a strong relationship with students outcome, except in the beginning. Providing some evidence that teachers in their first year ‘learn by doing’. The evidence on teachers certification is mixed but positive results are small. Overall most observable characteristics seem to have little impact on students results.

Besides these two review studies there have been several new studies that try to discover which teacher characteristics matter for students achievements.

Aaronson et al. (2007) use a longitudinal file that links Chicago public high school math teachers with students achievement in specific classrooms.[5] Furthermore, a large range of other characteristics were present in the data set making it possible to control for a wide range of factors.

There first conclusion is that teachers do matter. They find that an one standard deviation, one semester improvement in a math teachers quality raises students math scores by 0.13 grade equivalent. Thus, in one year an one standard deviation improvement in a match teachers quality translates into an increase in math achievements equal to 22 percent of the average gain.[6] A higher quality teacher especially increased performances of African American students and students with low or middle ranged scores.

Their second conclusion is that the vast majority of the variation in teacher effects is unexplained by easily observable teacher attributes. Tenure, advanced degrees and teaching certifications, which are the variables that determine compensation in Chicago, only explain 1 percent of the total variation. Combined with other characteristics a total of 10 percent of the total variation of estimated teacher quality is explained, leaving 90 percent unexplained.

In another study Kane et al. (2008) concentrate only on the difference between certified, alternatively certified and uncertified teachers. They looked at math and reading achievements of students in grades 4 through 8. They too find that teachers matter, and find that a one standard deviation increase in teacher quality increases students scores with 0.1 standard deviation in math and reading scores. Which is of similar magnitude than Aaronson, Barrow and Sander (2007), who's results translate to a gain of 0,15 standard deviation in math scores. More surprisingly, although there are large variations between teachers effectiveness it does not seem to matter if a teacher is certified or not. They conclude that the emphasis on certification status of teachers may be misplaced. However, this research, and previous mentioned research in Hanushek and Rivkin (2006) and Wayne and Youngs (2003) is based

[5] They specifically concentrated on 9th grade math achievements. Students in 9th grade are typically around 14 to 15 years old.

[6] A grade equivalent is a national normalization that assigns grade levels to test scores. For instance, a 9.2 implies that the student performance at the level of a typical student in the second month of ninth-grade. Hence, a typical student is expected to rise 1.2 grade equivalent in a given year. The research shows that a one standard deviation increase in the quality of a teacher increases the grade equivalent with an additional 0.26 in one year. Furthermore they find an average gain of 1.15 grade equivalents. In other words an increase of around 22% $(\frac{0,26}{1.15})$

on US data. How this exactly compares to teacher certification in the Netherlands is not clear, thus some caution should be exercised in generalizing these results to the Netherlands.

Rivkin et al.(2005) further add to this literature by focusing on academic achievements. They have a data set that contains date of students in grades 3 through 7. They also conclude that teachers matter and find similar results as the studies above. Yet, similar to the other literature, they do not find any significant evidence that having a master's degree improves teacher skills. They also conclude that teachers quality increased with the first years of experience. Yet, after three years of experience no significant results are found anymore.

To conclude, teachers matter, a lot. Yet, contrary to what many believe, high quality teachers are not identified by experience, academic achievements or certification. Nonetheless, these are exactly the criteria on which teachers are rewarded. The current system does not help to retain or attract top quality teachers, which is so important. Moreover, in chapter 2 it became clear that teachers them self are also not so happy about their salary. Which begs the question if the current system can be improved, hence the discussion about the introduction of PRP.

Chapter 4: Theory of Performance-Related Pay

"Call it what you will, incentives are what get people to work harder. "

- Nikita Khrushchev

The cornerstone of modern day micro economics is that people respond to incentives. It is one of the driving forces behind the 'personnel economics' literature. In this section I will go over the basis theory of performance-related pay and explain the two main advantages of performance-related pay over fixed pay. Namely, incentives and sorting and selection effects. I will then use the four principles of performance-related pay to elaborate the basic theory. The four principles will touch upon issues of measurement, multiple tasks, monitoring and including multiple information sources into the contract. To conclude this chapter I will also address some other issues that are not covered by the four principles, but are still relevant to discuss. The additional subjects include team pay, relative pay and distorted behavior. In chapter 5 I will elaborate on possible issues that arise because of interactions with psychological factors.

4.1. Basic Theory

When we think about compensation we can roughly define two kind of payments (Lazear, 1986). Namely, payment on input and payment on output. Payment on input, that is payment of a fixed salary, is based on the amount of skills or time worked. For example a teacher whose annual salary is specified in the beginning of the year, and at the end of the year is paid exactly that amount, independent of his output. Given of course that his performance does not fall below some threshold, i.e. if the teacher does not show up for work he might be fired. On the other hand, we have payment on output, which usually refers to payment on the basis of a measurement of performance (often called performance-related pay (PRP)). Take for example a fruit picker who is paid per kilogram harvested. Nowadays teachers are mainly paid on the basis of input. As we saw in chapter 2 teachers salary in the Netherlands is based on time worked, experience and their academic achievements. Unfortunately, these input factors seem to have little effect on students performance. Resulting in suboptimal outcomes. In theory output based pay has two main advantages over input based pay (Lazear, 2003).

4.1.1. Incentives

The first advantage is that output based pay creates incentives. Rewarding teachers (or schools) on an agreed performance measure will align incentives directed at teachers with those directed at students and potentially of society as a whole. Aligning incentives may create more efficient behavior via two similar ways. First, if for example wages are based on students performance, it creates a signal towards teachers what is important and what is not. If these signals are absent teacher might emphasize material that is obsolete or no longer valued by parents or future employers. Take for example a math teacher who still does every calculation by hand instead of introducing students to graphical calculators and computer programmes.

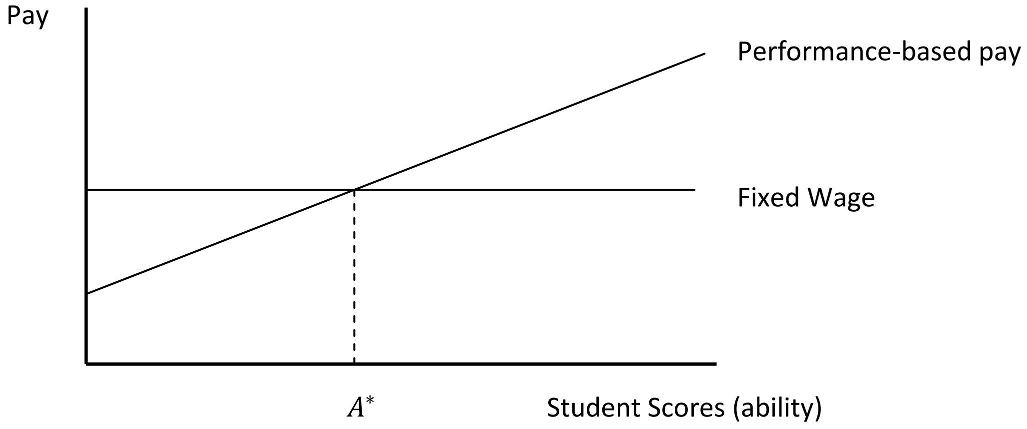
Second, it closes the gap that might exist between a teachers preference and those of her students. A teacher might fail to assign an additional assignment or to put in that little bit of extra effort on improving his lecture slides because it is too time consuming or tedious work. Even though he knows that assigning an assignment or improving his lecture slides would improve the results of his students. Compensating based on performance provides some incentive to put in that extra bit of effort and do the ‘right thing’.

4.1.2. Sorting and Selection

The second efficiency advantage of output based pay is that it may have sorting and selection effects. Assuming that a correct metric for teachers productivity is found (let’s assume for now that test scores of students is such a metric) and is used in the compensation scheme for teachers. Than teachers who are most able to raise test scores will benefit relative to those who are less able to raise test scores. Thus, attracting ‘good’ teachers while discouraging ‘bad’ teachers.

This can be very easily demonstrated in figure 4.1. In the figure you see an input based scheme, the fixed wage, and an output based scheme, a scheme that increases with students performance. The teachers that can raise the student scores above A^* earn more in the performance-based pay scheme than under a fixed salary. Hence, teachers of high ability are favored relative to low ability teachers. Subsequently, in this example teachers who are unable to raise student scores above A^* will earn less in a performance-based pay than under a fixed wage. Which will discourage ‘bad’ teachers to stay in the profession.

Figure 4.1: Fixed pay versus Performance based pay. From Lazear (2003)



Lazear (2000) provides empirical evidence for these two effects in the private sector. In his paper he studies a workforce of glass installers of the Safelite Glass Corporation. When the firm switched compensation schemes, from a hourly wage to a piece-rate scheme, there was a productivity gain of 44 percent. Halve of the effect was due incentive effects, the other halve was due sorting and selection effects. In fairness, figure 4.1 and the private sector example of Lazear (2000) most likely overstate the selection and incentive effects for a teacher incentive program. For one, figure 4.1 seems to implicate that the wage of teachers under performance pay would be largely variable, like a piece-rate scheme. We will see below (section 4.2), and throughout this theses, that in reality this is most likely not possible and even unwanted. The performance part for teachers can only be limited. It will therefore more likely be in the form of an end year bonus for top-performing teachers.

However, the study of Lazear (2000) does provides the evidence that these effect are not just theory and may give substantial efficiency gains. Obviously, the Safelite Glass Corporation is an almost perfect fit to test the theory. In practice, especially in the public sector, organizations may not fit the theory so nicely. Resulting in many complications when introducing performance-related pay. This will be the subject of the next two sections.

4.2. Principles of Performance-Related Pay

There may be several complications when using incentives in a contract. The agent might face multiple tasks or the output may be hard to measure. A good beginning is to look at the four principles of incentive pay that have been formulated by economist. Each deals with some of the issues that may arise when introducing incentive pay.

4.2.1. Measurement / Incentive Intensity Principle

The first problem that may arise is that effort is not directly measurable. In the case of a salesman or the glass installers above this might not be such a problem, as the number of sales or the amount of glasses that were installed provide a good proxy for the amount of effort that has been put it. But for teachers many argue that the value of services that teachers offer cannot be easily measured. First off all data has to be available, schools have to be able to link students to teachers. Often schools do not have such elaborate data records of their students and teachers. Second, the specific effects of the teacher has to be filtered out. For example, it is well known that test scores are influenced by the students socio-economic background. Rewarding on basis of test scores without taking this into account may leave the school with inaccurate results. Which will result in reduced motivation as teachers believe they are rewarded on factors beyond their control. In chapter 6 I will look at how the performance of teachers may be measured. The incentive intensity principle states how different factors influence the optimal share of incentives in a contract (Milgrom & Roberts, 1992, p221).

The incentive intensity principle assumes that the principle cannot observe the effort (e) of a risk-adverse agent directly, as in the case of teachers. However, it can observe another variable (z), for example students test scores. Yet, besides effort also a stochastic variable (x) determines the value of z. The variable x is for example the innate ability of a student or the social economic background of a student. A student test score is thus a combination of the effort of the teacher and the innate ability of the student, or:

$$z = e + x$$

As effort is not observable it cannot be used in a contract. On the other hand z (student test scores) are observable and thus can be used in a contract for the agent. Yet this induces risk on the agent as his payment is now also dependent on x, which he cannot influence. Thus, high effort of the teacher may be neutralized by a low innate ability of a student, whereas low effort of the teacher may be disguised by a high innate ability of a student. Suppose that a principal will only offer a linear payment structure to the agent then the wage of the agent will consist of a fixed part (α) and a variable part (β), or:

$$w =\propto +\beta z$$

The question which then arises is: what is the optimal β? Or in other word how large should the variable part be in the wage of the agent? The incentive intensity principle states the following[7]:

$$\beta = \frac{P^{'}(e)}{[1 + rVC"(e)]}$$

Were $P(e)$ is value of output for the principal, given effort level *e*. And $C(e)$ are the cost of effort for the agent, given effort level *e*.

It shows that β should be larger if:

1) the marginal benefits of effort on the value of output for the principal ($P^{'}(e)$) are high. In our case students/parents/schools probably find student achievements very important. Which would imply a high variable part.
2) The agent is less risk-averse (lower *r*). As the teacher profession is a rather stable job with lots of security. It might be argued that the current pool of teachers consist of relatively more risk-averse people. Who are less susceptible to a large variable part in their wage structure.
3) It is easier to measure the performance of the agent (lower V). As we will already have seen it is quite hard to measure the performance of a teacher. Thus implying a low variable part in the wage structure.
4) The agent is responsive to changes in β (higher $C"(e)$) i.e. if increasing effort does not have large cost for the agent. As we will see in some empirical experiments teachers are able to raise their effort level and thus are able to respond to a variable part in their pay.

4.2.2. Multiple tasks / Equal Compensation Principle

The second problem that arises is in the situation when the agent has to perform more than one task. In the basic theory it is implicitly assumed that the agent only performs one task. However, in reality agents often have to perform a wide range of tasks. These task are often substitutes of each other, as effort put into one task cannot be put into another task and vice versa. Teachers for example have to teach basic skills of math, reading and science; prepare students for exams; provide a good learning environment, etc.

[7] For a formal derivation see for example Appendix 6.2 of Economics and Management of Organizations (2003) by George Hendrikse.

Performing multiple tasks may complicate the design for a PRP scheme significantly, as incentives may work too well. Suppose that for task A the agent is paid a fixed salary while for task B the agent is paid a piece rate salary. It is no surprise that the agent will put all effort into task B, as this will increase his pay, while neglecting task A. This will hold for any combination of schemes as long as the marginal return to the agent for one task is higher than for the other. Ideally, the marginal return across tasks should be equal, hence the name equal compensation principle (Milgrom & Roberts, 1992, p228). This implicates that if the performance on one task is hard to measure (high V), and thus would receive little incentives, the strength of incentives for all tasks would have to be small.

In the case of PRP for teachers, that would tie students performance to teachers pay, a main concern is that teacher will focus heavily on test results of students. This 'teaching to the test' effect is unwanted as the teacher may fail to give students a deeper understanding of the material. The ultimate goal of education is critical thinking not achieving high test scores. In other words, student test scores may not capture the full range of tasks that a teacher has to perform. Rewarding on basis of test scores may thus results that teachers start to neglect some of his tasks. Jacob (2005) provides empirical evidence by studying high stake testing in Chicago public schools. He finds that on high stake tests the test scores of students significantly improved. However, on lower-stake tests a same improvement is not observed. We will see in chapter 8 that some papers studying PRP programs try to discover if the effects of PRP programs are caused by true learning or by teaching to the test techniques.

4.2.3. Other Principles

There are two other principles, the monitoring principle and the informativeness principle (Milgrom & Roberts, 1992, p219 and p226). The monitoring principle states the following: when an agents pay is (very) sensitive to performance (β high), it will pay off to measure that performance carefully (low V). Which explains all the effort of economists to invent better performance measures for teachers as this may allow higher variable pay. Finally, the informativess principle states that every piece of information that is correlated with the effort, but is not controlled by the agent, should be included in the contract. For example, contracts may also include subjective performance measures (assessments of principals) besides objective ones (test scores).

To summarize the above principles. First, the incentive intensity principle learned us that variable pay might be a good idea as the output of a teacher is very important. Yet, it might be difficult to introduce variable pay as effort is hard to measure. Second, matters are further complicated because teachers perform multiple tasks. The equal compensation principle tells us that if each task is not awarded the same we should be cautious, as unwanted behavior may arise. Third, the monitoring principle tells us that if we want to introduce variable pay there should be a good information system to measure teachers performance. And last, the informativeness principle states it might be a good idea to also include others measures of performance in the contract. Overall we can conclude, while introducing performance-related pay may improve efficiency, the magnitude in which it can be used for teachers pay is limited.

4.3 Other issues

Yet, besides these principles there may be further complications when introducing PRP. Teachers may work in teams, could also be awarded on a relative basis or may try to cheat.

4.3.1. Team work

Until now we only have dealt with individual pay, yet teachers may work, to considerable extent, in teams. If teachers are nonetheless paid on individual performance, it is feared that they will reduce cooperation with other teachers. This may in the end hurt overall school performance instead of foster it. In a way we have a multitask problem as above, between effort put in individual tasks and effort put in teamwork. If teachers are paid on individual basis teamwork is not rewarded and will be neglected while they will focus on their individual tasks. Paying for teamwork has thus the advantage that it will enhance cooperation between teachers. A second advantage is that overall school performance may be easier to determine than individual teachers performance, as less data is needed.

Yet, there is also an important disadvantage of team pay and that is the fear of ‘free riding’. An agents paid on the basis of team output does not reap the full benefits of his effort, as it has to be shared among the team members. Agents thus might reduce their effort in hope to ’free ride’ on the effort of their team members. Resulting in suboptimal outcomes. Furthermore, team pay should only be used if work is indeed joint. Otherwise, it may only induce more risk on the agent, as he only has a limited amount of control on the choices of his co-worker. As teachers are an intermediate case, as many other professions, principals face a trade-off. Provide team pay that accommodates teamwork but has reduced incentives, or provide individual pay that increases incentives but may undermines teamwork. Of course a combination of the two is also a possibility. Ultimately, this is (again) an empirical question. Unfortunately, it will become apparent in chapter 7 that the empirical evidence does not provide much help in deciding which one is better in the case of teachers.

4.3.2. Relative pay

Before continuing I also want to discuss one other way to possibly introduce PRP. Until now we have implicitly assumed that the teachers were rewarded on basis of some absolute standard. For example the reward may be based on some piece rate like scheme and every teacher gets rewarded accordantly. Yet another method to create incentives is via relative pay, i.e. via a tournament structure first introduced by Lazear and Rosen (1981). In this kind of scheme agents will have to compete with each other and only the top performers will earn an additional wage or earn a certain promotion (quotas). Rewarding via a tournament structure provides several advantages and disadvantages.

The first advantage is that often less data is needed. The principal does not have to know the exact performance of teachers, he only has to determine that one teacher is better than the other. Second it reduces common risk, as everybody faces the common risk it does not influence the relative performance. In contrast, common risk does influence the absolute performance. For example, in a given year the group of students are extremely bad (the common shock). For the relative ranking this makes no difference as all teachers face the same group of bad students. Yet the absolute performance of all teachers decreases as the students are just not capable of learning as much as students from previous years, reducing the wage of all teachers.[8] However, imagine the above example but with a tournament among schools. If one school just happen to have a particular bad cohort of students, while another school has a particular good cohort of students (school specific shock). The tournament will differentiate between lucky and unlucky schools instead of good and bad schools. Kane and Staiger (2002) show that this 'sampling variation' can have large impacts on the ranking of schools. So some caution is necessary when using relative pay. To summaries the above, in theory a tournament is better when common shocks are more salient than specific shocks (Green & Stokey, 1983). Which of the two shocks is greater depends on the situation and will have to be determined on an individual basis. Third, it may be a cheaper way of providing

[8] Obviously this assumes that all teachers are equally capable to handle 'bad' and 'good' students, which might not be the case.

incentives as only a limited amount of people have to be awarded. In other schemes every agent/teacher has to be awarded. But there also some disadvantages. First, agents may undertake costly actions that undermine the performance of other agents, i.e. sabotage. Second, it may reduce cooperation as you compete with one another. An competitive environment will emerge instead of a cooperative one, which is not always desirable. Third, some agents may be demotivated as they do not believe they can win the tournament and reduce their effort. Fourth, effort withholding norms may arise against high performers.

It is hopeful to note that still adoptions and improvements are made to the original model of Lazear and Rosen that try to invent new ways of paying agent for their relative performance, see for an interesting case for PRP for teachers (Barlevy & Derek, 2009). Both relative as absolute schemes are introduced in to practice. We will take a look at them in the chapter 7.

4.3.3. Distorted behavior

The last issue I want to address is that PRP may induce distorted behavior among teachers and schools. Jacob and Levitt (2003) provide evidence that teachers downright cheated when grading test scores for students. In their data set of Chicago public schools they found, trough the use of algorithms, that 4 to 5 percent of the classrooms cheated each year. Yet, as humans are quite creative, cheating is only one of many ways to game the system. For example schools may try to relocate bad students in special education or suspend them on test dates (Figlio & Getzler, 2002). When introducing and designing performance related pay policymakers should be aware of the possibility of distorted behavior and should implement safeguards. Although completely safeguarding the system might be impossible. Thus, when introducing and designing a performance related pay scheme it has to be taken into account that cost will be made safeguarding the system and some efficiency will be lost by an amount of distorted behavior by the agents.

Chapter 5: Psychology and Performance-Related Pay

Further complications may come from interactions between performance pay and psychological factors. A well known concept is that of intrinsic motivation and its possible 'crowing out' effect of explicit incentives. Raising the question if PRP should be used at all. In this section I will go over a few different psychological concepts and their meaning for performance related pay. Beginning with intrinsic motivation followed by reciprocity, social norms, status and fairness. As intrinsic motivation is often (mis)used as an argument against performance-related pay I will take considerable time to discuss it.

5.1 Intrinsic Motivation

Nobody will argue against the fact that people enjoy some tasks more than other tasks. People may derive direct pleasure by engaging in an activity or task. For example in surveys teachers state that they have become a teacher because they enjoy seeing improvements in pupils (Burgess et al., 2001). Intrinsic motivation opposes general economic theory which assumes that agents dislike to exert effort. Or more formally, effort is associated with negative (marginal) utility at all levels of the activity.

This would not be much of a problem if intrinsic motivation and extrinsic incentives would not interact with each other. As giving incentives would then have no effect on intrinsic motivation. Yet, in the social psychology literature there is mention of a 'crowding out' effect. It states that the introduction of explicit monetary incentives decreases task-specific intrinsic motivation. One of the consequences would be that putting incentives into place may decrease intrinsic motivation and hence effort. There are two theories explaining how this crowing out effect may work. Self perception theory and cognitive evaluation theory (CET) (Fehr & Falk, 2001).

Self perception theory assumes that people do not have perfect knowledge about the reasons why they perform a task. Which means that they do not exactly know to which extend intrinsic motivation determines their behavior. To assess the reasons for performing a task they look at the circumstances under which they perform the task. For example, if there are strong monetary incentives, which would cause the individual to perform the task regardless of his preferences, it is likely that individual will infer that his behavior is extrinsically motivated. If, on the other hand, an individual undertakes a task even when there are no external incentives to do so, he likely infers that his behavior is intrinsically motivated. Self perception theory is thus a theory of self attribution of motives (Fehr & Falk, 2001). The most interesting case is when a task has external incentives but at the same time is also intrinsic rewarding. Such a specific task is called an over sufficiently justified task by social psychologists. They claim that because external incentives are specific and prominent, while intrinsic motivation is more vague and uncertain, the individual will attribute the performance on an over sufficiently justified task to the external incentives. Hence, intrinsic motivation is supposedly crowded out.

Cognitive evaluation theory (CET) has a similar messages, but with a little nuance. It assumes that people have need for self-determination and competence. The effect of external incentives on intrinsic motivation can therefore be twofold. If external incentives are seen as controlling, reducing self-determination, intrinsic motivation will decrease. In contrast, if external incentives increase the feeling of competence, intrinsic motivation may increase. One of the hypothesis is that tangible rewards that are based on performing, engaging or completing a task are likely to be considered controlling and reduce intrinsic motivation. Thus, both theories predict a crowing out effect of intrinsic motivation when external *monetary* incentives are used.

A classic experimental study in this literature is that of Deci (1971), but many more followed. These experiments normally consisted of three phases, with a control and treatment group. In the experiments participants were asked to solve some kind of puzzle but, if they liked, could also read magazines. In phase 1 and 3 the subjects did not got paid for solving the puzzles. In phase 2, however, subject in the treatment condition got paid while the control group got paid nothing. The time spend on solving puzzles was taken as a measure of intrinsic motivation. Furthermore, a self reported measure of intrinsic motivation was obtained through a survey.

Different meta-studies were undertaken to gain additional insight in the experimental findings, which were done mostly on children and high school students. Arguable, the two newest and most elaborate meta-analyses are of Eisenberger et al. (1996) and Deci et al. (1999) both with complete opposite conclusions. Eisenberger et al. (1996) write in their conclusion that in the papers they have analyzed there is no sufficient proof to conclude that rewards reduce intrinsic motivation. In contrast, Deci et al. (1999) write that the 128 experiment they analyzed leads to the conclusion that monetary reward substantially decreases intrinsic motivation. See for summary statistics of both meta-analysis Appendix A1 and A2.

Unfortunately, the study of Eisenberger et.al (1996) had several methodological issues greatly reducing its value. For example: it misclassified several studies; it did not include some studies published during the period they covered; it omitted 20 percent of studies as outliers; it included studies who had inappropriate control groups, et cetera. For a further discussion of the issues concerning the meta-analysis of Eisenberger et al. (1996) see Deci et al. (1999) and the papers cited therein.[9]

The study of Deci et al. (1999) did not suffer from these methodological issues. The results that were found in the meta-analysis provided evidence for CET. As when verbal rewards were used, an external incentive that increases the feeling of competence, intrinsic motivation increased. To the contrary, when monetary rewards were used, an external incentive predicted to be perceived as controlling, intrinsic motivation decreased. This results was further confirmed by the finding that in the case of controlling verbal rewards, intrinsic motivation also decreased.[10] So at first sight intrinsic motivation seems to be an important factor to take into account. It evens raises the question if you should implement monetary reward at all if agents have high intrinsic motivation, which may be the case for teachers.

5.2 Criticism on Intrinsic Motivation

However, some economist are not convinced of the relevance of intrinsic motivation for economic theory, and have critically examined the evidence of intrinsic motivation and it effect on monetary incentives. They have problems both with the empirical evidence as well as with the theoretical concept.

Kunz and Pfaff (2002) have a critical look at the meta-analyses of Deci et al. (1999). They highlight the fact that, in contrast with free choice behavior, the self-reported measure does not provide a significant results if all studies are combined or if its limited to studies were the reward is dependent on performance. Deci et al. (1999) try to explain this by arguing that the self report measure is actually not really a good way to measure intrinsic motivation. Their main argument is that people might confuse their interest in the task with the enjoyment of receiving a reward, inflating the reported measure of intrinsic motivation. Yet in the past the same authors have written a paper that the best way to determine intrinsic motivation is to measure both free choice behavior and self report measures. The free choice measure was also

[9] The discussion starts at page 663.

[10] An example of a controlling verbal reward is: “just as you should”. Were the word ‘should’ is suppose to signal a controlling statement.

deemed imperfect as it was not able to differentiate between intrinsic motivation and ego-involvement. Only if both measures are correlated within conditions and studies it can be considered as intrinsic motivation (Deci et al., 1991). Making the argument not very convincing.

Kunz and Pfaff (2002) further note that a clear aging effect was found in the meta-analyses. As Deci et al. (1999) themselves write: “tangible rewards are more detrimental for children than for college students on both measures” (p.656). Greater cognitive capacity of college students and being more accustomed to the use of rewards are pointed out as possible explanations. College students may be more likely to interpret rewards as indicators of their effective performance rather than controllers for their behavior. This provides a clear warning of generalizing these results to a workplace environment as employees are even more accustomed to receiving monetary rewards for their performance. At least it points toward a limitation of the detrimental effects of monetary incentives on intrinsic motivation when used in an real organization.

Eisenberger et al. (1999) follow up this last point. They agree with the core assumptions of CET but argue that pay for performance is not perceived as controlling, and thus does not negatively relate with self-determination and intrinsic motivation. Reversing the ‘original’ negative relationship between monetary rewards and perceived self determination to a positive one. Eisenberger et al. (1999) conducted three experiments to invigorate their argument.

In experiment one 435 college students were asked to solve a number of ‘find the differences’ games. This time around Eisenberger et al. (1999) take a slightly different approach to measuring intrinsic motivation. This experiment consisted of two phases. In phase 1 the participants had to solve the games, and depended on if they were in the control or treatment group, got paid. In phase 2 the participants got 5 min of free time while the experimenter supposedly needed to get a questionnaire. The participants got to choice to solve some more puzzles or do something else. Time spend on the puzzles in period of free time is taken as measure of intrinsic motivation. They found that when a reward was offered there where positive effects on free time spend, perceived self-determination, perceived competence and task enjoyment. Implying a positive effect of explicit monetary incentives on intrinsic motivation.

In experiment two there was a survey administered under 348 employees of a certain organization. Employees with strong performance-reward expectancies showed higher perception of self-determination. The higher perceived self determination subsequently caused higher perception of autonomy, which, in turn was positively related to employees belief that the organization valued their well-being. The employees who experienced high autonomy, stemming from performance-reward expectancy, reported that they felt more active, enthusiastic, and energetic on a typical day at work (Eisenberger et al., 1999).

In experiment three there was again a survey administered, now under 367 employees of an organization. Employees who had a higher performance-award expectancy had higher expressions of interest in their work. This relationship was greater among employees who had a higher desire for control.

Thus, the core assumption of CET that there is a positive relation between self determination, competence and intrinsic motivation still stands. Yet, in these series of experiments it is found that monetary incentives do not have a negative relation on self determination as predicted by CET but a positive one. Hence, raising intrinsic motivation rather than decreasing it.

Fehr and Falk (2001) point towards further limitations of empirical evidence on intrinsic motivation. First, they are worried about the setup of the experiments. They claim that the studies cannot differentiate between a possible disappointment effect, that may arise when the payment gets removed in phase 3, and deterred intrinsic motivation. Making the results unreliable and biased toward finding detrimental effects. Second, as the empirical evidence does not contain field studies it faces the general limitations of experimental lab studies. For example, solving a maze in a lab study may induce very different behavior than tasks normally done at work, i.e. it lacks generalizability. Third, they note that even if there are negative effects of monetary incentive it is the total sum of incentives that counts. Thus as long as positive effects of introducing monetary incentives outweigh the negative effect of reduced intrinsic motivation it will be efficient to introduce them. Unfortunately, the psychological literature does not control for the total returns from a subjects performance, making it impossible to determine the total efficiency when monetary incentives are introduced.

Besides the shortcomings of the empirical evidence there is also criticism on the theoretical concept of intrinsic motivation. Kunz and Pfaff (2002) have three main problems with the theoretical concept. First, they find the terminology confusing. There is no precise distinction between intrinsic and extrinsic motivation, and it lumps together quite distinct issues (like internal drive, behavior, behavior of others, etc.). The terminology also seems to imply that motivation is mutually exclusive as it come either from within or from outside a person. Yet, there might be complex interaction between the two, determining human motivation jointly. Or as Johnson (1986) quite nicely wrote: “To say that teachers are motivated *primarily* by intrinsic rewards does not necessarily mean that they are motivated *solely* by them”. They conclude that the overall terminology is rather sketchy and vague.

Second, they argue that the precise mechanism through which intrinsic and extrinsic motivation works are still rather unclear, and far more complex than previously thought. Third, the theory does not provide help in identifying under which conditions intrinsic and extrinsic motivation are at stake. For example, the CET states that when something is perceived as controlling intrinsic motivation will diminish. Yet it does not provide objective criteria to determine when something is considered controlling and result thus rest on intuitive and post-hoc classifications. The studies of Eisenberger et al. (1999) nicely prove this point by finding that monetary incentives are not perceived as controlling by their subjects. Fehr and Falk (2001) give another good example. They note that since in most cases people already receive some monetary reward for their work, the interesting question is not whether one should pay a reward or not, but in which form the individual should be compensated. Is this via a fixed wage, a piece rate, or via a bonus? Or should the organization provide a combination of all of the above? Sadly, the theory and empirics do not provide help to answer any of these questions.

To recuperate, at first sight intrinsic motivation may seem to throw a spanning in the works as it may make monetary incentives ineffective. But at closer inspection it seems that the theoretical concepts is not very clear. Furthermore, the empirical evidence of a crowding out effect is rather scant. Some research even point towards a positive relationship between intrinsic motivation and monetary incentives. Making intrinsic motivation not that of an issue as previously thought. To make matters very clear note that I do not argue that teachers are not intrinsically motivated. Instead I argue that monetary incentive do not have a negative impact on the intrinsic motivation of teachers. Nonetheless, researchers should always be careful of possible hidden cost that rewards may bring. As all kind of interactions may exist between psychological motives and monetary incentives. I will go over some of them below.

5.2 Reciprocity

Reciprocity is the notion that people respond kind or positive to kind actions, and will respond hostile or negative to hostile actions.[11] In this literature there is support for the idea that the agents perceptions of explicit incentives matter. If the agent perceives the explicit incentives as kind they respond with substantial higher effort than if the explicit incentive is perceived as hostile. How the agent perceives the incentives depends on the reference point, which can be manipulated by the framing of the incentives (Fehr & Falk, 2001). In a sense this is a bit the same as the concept of intrinsic motivation. Although now we speak in terms of hostile instead of controlling and kind instead of increased competence. This brings us to back to the criticism on the theoretical concept of intrinsic motivation. As part of the effects can also be explained by an already existing concepts like reciprocity, or effects that are caused by reciprocity are interpreted as intrinsic motivation. Do note that the two concept differ greatly in their explanation. Were intrinsic motivation refers to task specific factors and effort is increases (decreases) because an agent likes (dislikes) a task. Reciprocity refers to interpersonal relations were effort increases (decreases) because an agent perceives the action of the principal, introducing incentives, as kind (hostile). Reciprocity thus has nothing to do with the task the agent performs. However, much of the evidence is based on experimental evidence and does not cover situations were principals want to introduce a variable part to an already existing fixed wage. Comparable to the experiments from intrinsic motivation. Therefore it is hard to determine what the exact effects will be of the interaction between reciprocity and explicit incentives in the case of performance related pay for teachers. The only helpful thing that policy makers can take from current literature of reciprocity is that framing the incentives positively seems to be more effort enhancing than framing them negatively.

5.3 Social Norms

Social norms are rules of behavior that coordinate our interactions with others. In combination with explicit incentives either effort-enhancing or effort-withholding norms may arise. For example, paying on basis of team output may induce people to develop effort-enhancing norms because if one team member shirks it will reduce the payment for all team members.

[11] The Norse Vikings from the Viking age nicely illustrate the idea of reciprocity in one of the verses from Hávamál ('Sayings of the high one'). Attributed to Odin it offered advice for living, proper conduct and wisdom. In one of the verses it states: "To his friend a man should be a friend and repay gifts with gifts; Laughter a man should give for laughter and repay treachery with lies".

Shirkers thus face cost of disapproval by the other team members. This threat is very real as evidence suggest that people will punish even if this implies considerable personal cost (Fehr & Gächter, 2000). In contrast, in tournament structures effort-withholding norms may arise. High performers form a negative externality to other competing workers as the higher performance of others reduces the chance of winning for a given agent. This may also be the case in a piece rate system that is subject to the ratchet effect. High effort is beneficial to the worker but also increased the probability that the firm will adjust the piece rate scheme in the future, again inducing a negative externality for other workers. In short whenever an action provides positive externalities to other agents effort-enhancing norms may arise, when actions provides negative externalities effort-withholding norms may arise.

5.4 Status / Recognition

Getting an award or bonus might also increase motivation by the implied ‘thank you’ of the reward or as recognition of the work that has be done. Winning teachers and schools may also derive status of winning an award because they are ‘the best’ (Burgess et al., 2001).

5.5 Fairness

Maralidharan et al. (2009) point towards an interesting finding in India. Teachers that had the highest absent rates had the highest job satisfaction. In contrast, teachers that had the lowest absent rates had the lowest job satisfaction. Teachers who were absent often were quite happy as they only had to exert low effort and still kept their job. Yet hard working teachers got demotivated as there was no difference in pay between them and shirkers. In this particular case the lack of external reinforcement of performance decreased motivation. Fairness might thus be an important factor in performance related pay schemes.

These are only a few other concepts that might influence the way a PRP scheme may work. There may be many more psychological factors that interact with explicit incentives. These interactions may be surprisingly and unforeseen. Policy makers should be on the watch for these interactions as they may cause desirable behavior but, more importantly, may also cause behavior which is undesirable.

Chapter 6: Measuring the Performance of Teachers

"In teaching you cannot see the fruit of a day's work. It is invisible and remains so, maybe for twenty years".

- Jacques Barzun

In this section I will go into more detail how teachers performance might be measured. We have seen in chapter 3 that input factors do not necessarily indicate good teachers performance. Paying on basis on output is the alternative but introduces all kind of new problems. As the quote of Jacques Barzun demonstrates the true output of a teachers might be rather complex to measure. The output of a teacher is not only getting students to achieve good grades it also entails learning student to critical think, encourage their creativity and getting them ready to participate in society. In this chapter I will discuss the advantages and disadvantages of the different objective and subjective criteria's on which teachers performances may be assessed.

6.1. Objective Criteria

Objective criteria's uses measures like test results and pass rates to assess the teachers performance. Measuring performance on objective criteria introduces three general problems. First, test scores are imperfect metrics to assess the knowledge of students. A test has limitations both in size as in how well it assesses the knowledge of the student. As we have seen in section 4.2.2. this may result in 'teaching to the test' techniques that will limit students true learning. Two, test scores do not contain information about the other tasks of the teacher, like encouraging creativity. Three, it might induce teachers to cheat. Besides these general problems there are also implementation problems of using test scores and pass rates.

6.1.1. Test Results

The most simple way to use test results is to just take the average school/course results of the student and base pay for the teacher on that. Unfortunately, this is often not a good idea. We already stated that test scores contain lots of outside influences that the teacher cannot control. For example, the social economic background, current level of knowledge, school effects etc. Without taking these factors into account there will be hardly a fair and accurate view of the teachers performance.

Ladd (1996) gives a nice example of this point. South Carolina had an incentive program that used test scores but did not control for social economic factors. After some time in the program administrators realized that schools that served lots of students from low social economic backgrounds were always at the bottom, and never won awards. They tried to adept the program by classifying schools into five different groups, and the top 25 percent of each group that showed the greatest improvements won awards. Yet this provided another problem as schools who were at the bottom of a group claimed that they were classified too highly. Subsequently, schools tried to influence in which groups they were classified. This example illustrates that making things simple is not always the best solution.

6.1.2. Pass Rates

With pass rates, or similarly some threshold that needs to be met, another problem arises. Assessing teachers performance on pass rates may introduce ‘threshold effects’. Threshold effects arise when teachers only concentrate on students on the bubble, i.e. the students just below or above the threshold. This behavior is rational for the teachers as they are only awarded on the basis that students pass. Top performing students already pass or are above the threshold, while low performing students may be considered ‘hopeless’. For example, Neal et al. (2010) conclude that the No Child Left Behind Act (NCLB) of 2001, that used proficiency counts to asses teachers performance, may leave 25.000 children in Chicago public schools behind due the focus on children who are on the bubble. However, this does not mean that, for example, the NCLB was ineffective. It only states that it may not target the entire student population. Ironically, it may leave behind the most disadvantaged students.

6.1.3. Value-Added Models (VAM)

Value-added models may be the solution to these problems and allow administrators to use test scores and pass rates in a better way. Value-added models try to isolate the effects of a teacher or school by controlling for a wide range of factors. There are basically two versions. The first version requires a vertically scaled test such that the average gain can be calculated by subtracting the prior test scores from current test scores, while controlling for socio-economic characteristics. The second version uses a statistical model (regression), that controls again for a wide range of characteristics, to calculate what a student should score. To determine a teachers performance the actual results of the student is compared with the predicted score. The second method has the advantage that it does not need a vertically scaled test. The VAM thus bases the performance of teachers on the improvements students make while controlling for outside influences.

However, VAM also have several drawbacks. The first problem is that these methods can become quite complex. Although these statistical models might make a lot of sense to economists, teachers probably beg to differ. They will often have a hard time understanding how this value-added is calculated. Especially as the value-added score may be very different than the original test score. This may cause several complications. A teacher that performs bad will likely claim that performance measure is plain wrong or manipulated somehow. Causing distrust and unrest among teachers and against administrators. An illustrative quote from a Dallas teacher: "look, I don't know a CEI (the value added measure) from Yankee Doodle, but if you're going to grade my performance on faulty information, I have a problem with that" (Milanowski, 2008). Second, because teachers do not understand how the value-added is calculated they might also not understand how to improve their value-added score. We might speak of a tradeoff between fairness and clarity. While the VAM might be more fair as it controls for outside influences it may be less clear for teachers. Third, the VAM results may be biased. For example, VAM may omit or incorrectly model influences on students achievements, thus biasing the results. It may therefore only be reliable at identifying the top and worse performing teachers (McCaffrey et al., 2003), which is why VAM are often used in collaboration with tournament setups (Milanowski, 2008). McCaffrey et al. (2003) conclude in their book that the magnitude of the possible biases that influence VAM are unknown. Therefore the use of VAM in high stakes environments is not to be recommended. However, as McCaffrey et al. (2003) also note, other methods might be even more biased and VAM may be the best we got.

6.2. Subjective Performance Criteria

We have seen above that measuring the performance of teachers by objective performance measures may be rather difficult, as test scores may be noisy signals and may not reflect the full performance of the teacher. One solution to this may be the incorporation of subjective performance measures. As we have learned from the informativeness principle in section 4.2.3. including additional information about the performance of the agent may make the contract more efficient. Supervisors, in this case the school principals, are often able to subjectively asses the performance of a teacher. Which may complement and improve the objective measure as principals may pick up things that test scores do not and vice versa. The drawback is that ex-post haggling may arise because the criteria are not verifiable and therefore cannot be settled via legal matters. The agent thus has to trust the supervisor that he will assess him honestly. Furthermore, agents may undertake unproductive activities to improve evaluations by the principal.

The first question to ask is if supervisors of teachers can indeed reliable assess the quality of teachers. Jacob et al. (2005) tried to answer that question in their paper. They administered a survey under elementary school principals asking them to rate their teacher on several dimensions.[12] They then compared the results of their survey with a value-added measure of teacher quality they calculated. They found two main results. One, the correlation between how effective the teacher is at raising math and reading achievement and the value-added method is 0,32 and 0,36 respectively. This increased to 0,56 and 0,38 if average student achievement were used instead of the value-added measure (which tries to measures improvement rather than average results). Suggesting that principals base their rating at least partially on a recollection of students test scores in a particularly class. Jacob et al. (2005) also try to determine which factor predicts future student achievements the best. They regress math and reading results from 2003, controlling for different characteristics, on teacher experience, education, salary, principal ratings and the value-added measure. Only the principal rating and the value-added method have significant predicting power. Which let them conclude that principal ratings predict future student achievement better than teacher experience, education or salary, although the value-added measure is even better.

Two, the teachers that were identified by principals as being in the top category were in the top category according to the value-added measure about 52 percent of the time in reading and 69 percent of the time in mathematics. For the teachers at the bottom these percentage are 42 and 69 percent. For categories in the middle this drops to around 50 percent. This seem to suggest that principals have some ability to identify teacher at the top and bottom categories, but are less successful at the middle. Overall principals seems to have a decent ability to assess which teachers perform and which do not. Although, like VAM, it is also an imperfect measure of teachers performance.

The second interesting question is how large are the drawbacks of subjective performance measures. Jacobs et al. (2005) also provide some results on this. In the study

[12] Principals were asked to grade their teachers on a scale from 1 to 10. On the form several teacher characteristic were mentioned, like dedication and work ethic, student satisfaction and raising student math/reading achievements. The principal than had to grade the teacher on each characteristic, at the end they had to grade overall teacher effectiveness separately.

principals discriminate against male and untenured teachers, giving them lower ratings when scoring their ability to raise students achievements. Strangely, on the overall measure of teacher quality there is no discrimination. Making the lower scores for male and untenured teachers on their ability to raise student achievements a bit of an anomaly, with no clear explanation. Interestingly, teachers more highly rated on having a positive relationship with the administrator, were more highly rated by the principal. Which points out that possible non-productive influence activities of teachers might pay off.

A big drawback of this study is that no consequences were attached to the assessments. In a high stake environment, like assessing a teacher for a PRP scheme, behavior will likely change. As Podgursky and Springer (2006) note old programs in the US that uses subjective measured were abandoned because they were: “highly susceptible to gender and racial discrimination as well as nepotism (favoritism)”.

To summarize, having a reliable way of measuring teachers performance is the greatest challenge that PRP for teachers will have to face. Value-added models have come a long way in isolating the performance of teachers. However, VAM are still an imperfect measure and thus should not be used careless. Subjective measures may complement VAM but again are imperfect and may induce unwanted behavior, both from teachers as administrators. Both criteria only seem to be able to reliable identify the top and bottom of performing teachers. Therefore it seems that making use of relative pay is more suitable for the current available performance measures. Overall, this difficulty in measuring teachers output limits the magnitude of the performance-related part in total teachers pay.

Chapter 7: Empirical evidence

In recent years there have been a surge in PRP programs for teachers. For example, there have been new programs in the United States, England and Portugal. However there have been few studies evaluating these programs. Moreover, as the programs have a wide range of designs it is difficult to draw general conclusions. The studies that have be done can be divided into non-experimental designs, quasi-experimental designs, randomized designs and field studies (Shadish et al., 2002). In this chapter I will describe each study in detail. I will describe the PRP programs, the methodology used and will give commentary on each study. At the end I will summarize the findings. I suggest readers who are unfamiliar with the topic of causal interference to first read Appendix B, as this will help to better understand this chapter. In Appendix D the reader can find a table shortly summarizing all of the studies.

7.1. Non-experimental design

Non-experimental designs refer to designs in which a presumed cause and effect are identified but lack proper control. They simple try to observe the size and direction of the effect among variables but cannot draw any causal conclusions.

7.1.1. Figlio an Kenny (2007)

Figlo and Kenny (2006) provide a first look of the possible effects of PRP programs. They combine data from the National Education Longitudinal Survey with a survey of their own. They find that schools that answered that they used some kind of incentive plan raised test scores with 1 to 2 points. This happens only if the bonuses are awarded selectively, i.e. if a very large fraction of the teachers receive the bonus test scores did not increase. They also noted that the effects are stronger in schools that draw more students from low and middle income families.

However, Figlo and Kenny (2006) cannot report what causes this positive relationship between schools with incentive programs and student test scores. This could be indeed due the incentive program but it could as be due some underlying variable. For example, schools with incentive programs may be more likely to have innovative teaching methods. A causal relationship can only be established by (controlled) experiments.

7.1.2. Woessmann (2010)

Woessmann (2010) uses a cross-country comparison to see if countries with performance related pay score higher that countries without it. The Programme for International Student Assessment (PISA) is used to compare students results across countries. PISA is an international standardized test administered to a representative sample of 15-year-olds from each participating country. The PISA is held every three years with 190.000 students participating from 28 OECD countries. The OECD's International Indicators of Education Systems (INES) surveys was used to determine if a country had any kind of performance related pay. Unfortunately, this will not allow to discriminate between the scope and structures of the different incentive plans.

The main result is that student results are significantly higher in countries that make use of teachers performance pay. On average students in those countries score 25 percent of a standard deviation higher on the PISA math test. An interesting result is that in countries without performance related pay teachers with a tertiary degree in pedagogy have significant higher scoring students. However, in countries with performance related pay this difference disappears. Indicating that in countries with PRP teachers with lower educational degrees perform on average just as well as teacher with higher educational degrees. Different student characteristics do not seem to have any large significant effects. On the reading and science part of the PISA test similar results are found. Student in countries with performance-related pay score 29.9 and 24.1 percent higher, respectively. The results are robust to several estimation methods.

As before the study can only identify a general pattern between performance-related pay and students results and cannot identify any specific causes for this positive relationship.

7.2. Quasi experiment

A quasi-experimental design conduct an experiment in which the participant are not assigned randomly to the conditions. The researcher therefore has to deal with possible selection biases that may arise.

7.2.1. Elberts et al. (2002) – individual incentive program

Description

In 1996 a Michigan *alternative* high school introduced an incentive program. The main objective was to target student retention. The school had around 500 students who had problems succeeding in the traditional school system. These students often had attendance problems and/or were dropping out, hence the main objective of the program. Teachers were paid a base wage for each 60-minute class they taught. The incentive program supplemented this with two variable parts. One, a retention bonus was paid if 80 percent or more of the students were still enrolled and attending the course at the end of the quarter. This was determined by the principal, who took a surprise visit during the last week of the quarter. The retention bonus was approximately 12 percent of the base wage for teachers. Two, teachers who received an average rating of 4.65 or higher on student evaluations and maintained this for four consecutive quarters received an additional bonus. The student evaluations consisted of 15 items. This bonus raised their base wage with approximately 5 percent. An average teacher on a high school without bonuses would earn around $22.848. On the high school with the incentive program this could accrue to $27.412. The idea of the principals was that via increased retention rates student achievements would increase.

Methodology

The school that introduced the incentive plan was compared with a similar alternative high school in the area. As detailed student characteristics were not available the authors relied on administrators and district educators in selecting a suitable comparison school. An important difference between the two schools was that the incentive school was not part of an union while the control school was. The authors obtained five years of data (1994-95 through 1998-99). The data spanned two years before the program, the incentive year, and two years after the implementation of the program.

Elberts et al. (2002) used a difference-in-differences (DID) analysis to calculate the effect of the program. They looked at multiple student performance measures including course completion, class attendance, grade point average, and passing rates conditional on course completion. When a DID estimate is used the treatment and control group have to follow parallel trends. In this case this would mean that student characteristic, curriculum, local economics and other external factors would have to change similarly. Unfortunately, data limitations prevent the authors to formally test this assumption, limiting the internal validity of the study. Furthermore, as the experiment only encompassed one alternative high school, generalizability is low. Hence, this study should only be viewed as an useful first step in discovering the effect of performance related pay programs for teachers.

Results

When the program was implemented student retention rose from 50 percent to 70 percent . Yet daily attendance did not increase. Remember, for teachers to obtain the rewards students only needed to be present in the last week of the quarter, as that was the only time the principal would control the attendance. Moreover, grade point averages and pass rates in the incentive school decreased, leading to hypothesis that the incentive school retained low-achieving students, who where most likely to drop out otherwise. Anecdotic evidence further suggested that teachers changed their course content and instructional style to make their course more interesting and well liked by students. More field trips and in-class parties were observed.

Commentary

The above case illustrates that incentives may work, but, if badly designed, do not contribute to the ultimate goal. Teachers responded in a way that the principals had not anticipated. Teachers raised retention rates trough higher attendance in the last week of the quarter, instead of higher daily attendance throughout the period. This may have been easily countered by, for example, having multiple surprise visits during the period. Of course this would raise the (safeguarding) cost of the program. Teachers also managed to get more favorable student evaluations by making courses more fun. Unfortunately, but not completely surprisingly, both did little for students achievements.

The first step that Elberts et al. (2002) took provides an interesting lesson. The design of the performance related pay program is highly important!

7.2.2. Lavy (2009) - individual incentive program

Description

In 2000 the Israeli Ministry of Education initiated an experiment under forty-nine, poor performing, high schools that awarded individual teachers via a tournament structure. The experiment included all English, Hebrew, Arabic and mathematics teachers who taught classes in grades ten through twelve in advance of matriculation exams in these subject in June 2001.[13] The paper only evaluates English and math teachers as schools were allowed to replace the language (Hebrew and Arabic) teachers with other core subjects teachers.

[13] The Israeli system is somewhat similar to the Dutch system. The final matriculation score is a combination of internal (school-level exams) and external exams (national exams). If students pass enough courses they gain their matriculation certificate which gives them admission to university.

Teachers entered as many times as the numbers of classes he or she taught. The experiment was announced as a three year program but only lasted one year due budget cuts from the government.

The experiment was done as followed. The teachers were ranked on two fronts: the pass rate and students test scores. In both cases value-added methods were used. Thus via regression a predicted pass rate and test score was calculated and compared with actual scores. The value-added method included controls for social economic factors. Teachers were told explicitly that they were to be compared with other teachers of the same subject in the same school. Teachers with positive scores were ranked on the following basis:

Place	Test score	Pass rate	Final score	Prize ($)
1	16	20	30-36	7500
2	12	15	21-29	5750
3	8	10	10-20	3500
4	4	5	9	1750

The incentive program had the following notable design features. First, it employed an individual tournament setup. We have seen in chapter 3 that a setup like this may result in lower cooperation levels among teachers. Second, note that the same rank on pass rate is awarded with 25 percent more points than a similar rank on students test scores. Following the equal compensation principle teachers are expected to put in more effort into making students pass rather than increasing test scores, given that both require the same amount of effort. Weak students should benefit most from such a setup. Teachers that want to increase pass rates will focus on weak students as good students already pass. Third, the value added regression did not include lagged test scores. This would potentially open up a way for teachers to game the system. For example they might encourage students to perform badly in earlier test that are not included in the tournament. This would result in lower predicted scores and the teachers would show larger improvement among his students. The last notable thing of this study is that it conducted a follow-up survey among 74 percent of the teachers.

Methodology

Schools were selected for the program on the basis of three criteria. One, the school needed to be a comprehensive high school.[14] Two, matriculation rates for math and English were lower

[14] Secondary comprehensive school consist both of the middle and high school grades (grades 7-12) comparable to the Dutch 'Middelbare school'

than 70 percent in two instances over the last four year. Three, schools could only participate in the program if their overall 1999 matriculation rate was equal or lower than 45 percent (national mean). Ninety-nine schools met the first two criteria which Lavy denotes as the eligible schools. Forty-nine of these schools also met the last criteria, the treated sample.

Unfortunately, the treated and untreated schools in the eligible sample differed greatly from each other. The sample of treated schools was therefore not representative for the eligible sample. Making a simple comparison between the treated and untreated schools not possible. Lavy needed to find a different way to identify a comparison group. Luckily the matriculation rate that was used for the assignment of the schools into the program was not entirely accurate. The matriculation rate came from a preliminary and incomplete data file. This data file was later updated by the Ministry with corrected rates. Given the assumption that these errors in the data file are random an appropriate control and treatment group can be established. Explicitly, 18 of the 49 treated schools had (corrected) matriculation rates above the threshold, and thus should not have been part of the program. This is the treatment group. Each of these schools was paired with a school with an (almost) identical correct matriculation rate but who were not erroneously noted in the preliminary data file, the control group.

Lavy (2009) then uses a DID procedure to estimate treatment effects. Although Lavy cleverly managed to establish a treatment and control group it should be noted the design is not as strong as a randomized sample. First off all, it introduces an additional assumption that the measurement error in data file is random. Although Lavy makes a compelling case that this is indeed the case, he cannot fully test that some unobservable factors may influence the measurement errors. Second, as he uses a DID estimator his treatment and control groups have to satisfy the common trends assumption. As he only has one year of pre-program data this cannot be extensively tested. Third, his sample is reduced to only 18 schools. Decreasing explanatory power. To counter some of these concerns Lavy (2009) shows that his estimates are robust to different estimation procedures. Although there are some differences between the results of the different procedures. To summarize, the results of Lavy (2009) should be interpreted cautiously.

Results

The experiment produced the following results. The treatment group increased math pass rates 14 percent relative to the mean of the control group. For English the pass rates increased with

5 percent. Furthermore, average math scores increased by 10 percent while average scores for English increased by 4 percent. Test taking of math exams increased by 5 percent while English test taking increased by 4 percent. Lavy (2009) reports that for math the outcomes can be traced back to the first three quartiles, the just above and the below-average students. For English the effects originate from the first two quartiles. With the first quartile of students being the most important cause of improvement in both cases. A major concern is that these improvement are caused by teacher that are assigning higher grades for the internal exams, as this would increase their chances to win. Lavy (2009) provides evidence that this is not the case. To summarize, the experiment found positive results mostly originating from improvements by below average students.

To gain insight into how teachers achieved these results data was used from telephone interviews. The results indicate that teachers increased effort by adding special instruction time beyond their regular schedule, especially during the last few months of exam preparation. Math teachers reported that they taught more in small groups while English teachers tracked more by ability. English teachers also gave more attention to weak students. The author conclude that these results show that the cause of improvements were more likely from changes in effort en pedagogy rather than the introduction of 'teaching to the test' techniques. Sadly, the interview did not contain information about the attitude of teachers towards the program and possible changes in teacher cooperation levels, stress etc.

Commentary

Lavy (2009) find positive results for an individual incentive program that uses a tournament setup. The results are found in a educational system that is somewhat comparable with other Western educational systems. However, the studied sample does not reflect the full range of schools, as it only contains below average performing high schools. This becomes even more apparent in one of the alternative estimation methods that was used. Where including relative more schools with relative higher matriculation rates lowers the treatment estimates that were found. Also, as the program only lasted one year, long term effect cannot be established. Furthermore, the bonus payment were quite generous, the top performing teacher increased his salary up to 50 percent.

An interesting side note, Lavy (2008) also investigates in another paper if there is a difference between the performance between men and women in the program. No differences were found.

7.2.3. Atkinson, Burgess, and Croxson (2004) – individual based program

Description

Atkinson et al. (2004) evaluates the PRP program the UK government introduced in 1999. Prior to the PRP program teachers wage was determined by a single pay scale, with nine steps, ranging from £14.568 to £23.193 per year. Teachers usually progressed up in the scale in annual increments. In 1999 about 75 percent of the teachers had reached the top of the scale.

The reforms in 1999 changed this. Teachers who were at the top of the final scale could apply to pass the performance threshold and move up into the upper pay scale. In the upper pay scale teachers could gain additional increments, but these were tied to their performance. Besides moving up in the pay scale teachers would get an annual bonus of £2000 until the end of their career, it therefore had significant lifetime value. To pass the threshold teachers had to convince a head teacher that they had reached acceptable standards in five areas: knowledge and understanding of teaching, teaching management and assessment, wider professional effectiveness, professional characteristics, and pupil progress. On pupil progress teachers themselves had to provide evidence that their teaching improved their students results more than average. Passing the threshold was thus more about rewarding past performance, and cannot be considered true PRP. The PRP aspect of the reform was in the upper scale as the increments in the upper pay scale were tied to the performance of the teachers. Sadly, no further information was available about the effects of the PRP aspect in the upper pay scale at the time Atkinson et al. (2004) wrote their paper. In the end around 88 percent of the eligible teachers applied, and of these, 97 percent passed the threshold. Or as one teacher noted: “it is hard see who isn’t going beyond the threshold”. Although ex post the program may have acted more like a general pay raise, ex ante interviews with teachers revealed that teachers saw the program as a real incentive scheme.

Methodology

Atkinson et al. (2004) obtained five years of data. The data consisted of data from two years before the program, the introduction year of the reform and two years after the introduction of the reform. As most schools could not provide all the necessary data the authors required, the final sample was limited to only 18 schools. The 18 schools consisted of 182 teachers and around 23.000 students. The sample of schools that is studied is thus not representative for all schools. The sampled schools for example had better IT systems. To calculate the effects of the reform a DID analysis is used which compared the eligible teachers, who are at the end of the (first) pay scale, with the ineligible teachers. The two groups of teachers systematically differed between experience as the eligible teachers, by definition of the program, are more experienced. The authors are worried that the ability of teachers to improve students results may improve over time but at a decreasing rate. This could bias the results downwards as the ineligible group will show greater improvement because they are lower at the 'learning curve'. Therefore the authors also include a regression analysis that controls for teachers experience.

Results

To see the effect of the program on student test scores value-added measures were calculated on the basis of two standardized test held at the age of 14 and 16. The DID analysis shows positive signs for English and science teachers but negative signs for math. The overall DID estimate is positive but too small to be significant. Yet, as stated above, the DID analysis may underestimate the effects. The regression finds larger results and reports that the program added on average half a grade of value-added per child for eligible teachers, but again not for math teachers. The program also had most impact among low achievers, as we have seen in other experiments. The authors find the effect they found substantial as GCSE's exams are high stake exams that are the gateway to higher education. The question remains how valid these results are. It could be that the regression indeed picked up experience effects but it may also have picked up other factors.

Commentary

Atkinson et al. (2004) show mixed results. They find positive results for English and Science teachers but not for math teachers. Which may raise concerns for the general applicability of PRP programs for teachers. We also see that the design of the program is again important. Passing the threshold acted like a general pay rise but at the same time induced greater bureaucratic paper work. It therefore did not discriminate enough between teachers which will most likely limit the incentive effects, now and in the future. Moreover, it may point toward limitations of the subjective measures that were used. For future research it is interesting to see how the upper pay scale functions, as it has true component of PRP. Although data limitation might make it hard to find results.

7.2.4. Lavy (2002)– school based incentive program

Description

In 1995 the Israeli Ministry of Education also had initiated an experiment. The experiment contained sixty-two 'secondary comprehensive schools' and lasted two years. School performance was measured by three metrics: average number of credit units, proportion of students receiving a matriculation certificate and school dropout rates. For the performance measures a value-added measure was calculated, i.e. actual scores were compared with an expected base predicted from regressions. The value-added method included controls for social economic factors. The schools were ranked each year according to their improvements, the top third performers gained awards. School awards varied between $13.250 and $105.000. Schools had to distribute 75 percent to teachers, the rest could be spend on school facilities. Teachers on winning schools won between $250 and $1000 (mean annual salary is $30.000). The author notes that besides the monetary rewards the nonmonetary benefits, from publicity and improved reputation of winning schools, should not be overlooked.

Notable design features of this study are the following. First, the reader should note that the performance pay program does not directly link students test scores with pay but instead only uses measures that are related to pass rates. Second, it introduced a school based tournament setup. Only schools from small communities were allowed to participate, which may made the circumstances especially favorable for a school based program. Schools in small communities might have better social control against free riding, reducing possible negative effect from the (school based) incentive plan. Three, as they wanted to encourage schools to focus on weak students only the first 20 credit were counted in the measures to rank the schools. Four, it compared the program with an input based program.

Methodology

Schools were selected on the basis of two criteria. The first criteria was that the school was a secondary comprehensive school. The second criteria stated that the school should be the only one of its kind in the community.[15]

Lavy (2002) suffered from the same problem as Lavy (2009) as it did not have a proper control group in the original sample. This time he found out that schools that had another one of its kind in the community were comparable with schools that were the only

[15] Four kind of schools were identified: Secular Jewish secondary schools, boys and girls religious Jewish secondary schools and Arab secondary schools.

one of its kind in the community. As in Lavy (2009) he uses a DID procedure to estimate treatment effects. The common trends assumption is checked by looking at two years of pre-program data.

Results

Lavy (2002) found the following results. In non-religious schools the first year of the program had no significant effects. Yet, in the second year significant positive effects are found. Credit units increased by 0.7 units, average scores increased by 1.75 points and proportion of pupils taking the exam increased by 2.1 percent relative to the control schools. Pass rates (proportion of student entitled to a matriculation exam) did not significantly change. For religious schools positive effects are estimated in the first as well as in the second year, and are mostly bigger in magnitude. Furthermore, mean dropout rates were also reduced by around 0.5 percent. Weak students and students with poor social economic background seem to benefit the most from the program. Which seems to hint toward the idea that teachers responded to the design choice to only include the first 20 credits of each student. Lavy (2009) found similar results with another, but similar, design choice. Lastly, Lavy (2002) compares the cost effectiveness of this program with an input/extra resource program. Although the input program is more effective at raising halve of the achievements measures on the other halve the incentive program is more effective. Given that the input program costs twice as much, the incentive program is deemed more cost effective.

Commentary

Lavy (2002) find positive results for a school based, tournament style, incentive program. Interestingly, the program raises test scores without directly linking pay to it. As in Lavy (2008) it finds that weak student seem to benefit most from the program. As the program only studies schools in small communities the generalizability to larger communities poses a problem, especially as it is a school based program. Furthermore, as the program again lasted only a short amount of time, long term effects cannot be studied.

In my opinion Lavy (2002,2009) missed a great opportunity to help the debate of individual versus group incentives, as both had a similar designed incentive programs. Had the two studies reported on teachers cooperation levels, stress levels, level of free riding etc. interesting results may had surfaced. Also it would have been interesting if the interview had contained questions about how much teacher liked the program and if they thought it was fair.

Especially, since the two experiments were using VAM that may be complicated for teachers to understand and might be believed by teachers to be unfair.

7.2.5. Ladd (1999) and Ladd and Clotfelter (1996) – school based incentive program

Description

In Dallas an incentive system has been active since the school year of 1991-92. The paper studies the program until 1995. The Dallas incentive system covers all primary and secondary schools.[16] In each of the winning schools principals and teachers received a bonus of $1000 and other staff such as janitors and secretaries received bonuses of $500. Schools who where low in the rankings were more closely monitored but no further automatic sanctions were given. Twenty-eight percent of the schools won ones, 11 percent won twice and only 4 percent won three times. Which is unsurprisingly as schools are awarded for their improvement, which is hard to maintain for several years.

Schools were ranked on basis of students test scores of the TAAS and ITBS, both are standardized test. It uses a particular VAM method that uses a two-stage regression procedure to determine the ranking of schools. At the first stage the test scores are controlled for social economic factors. In the second stage this adjusted test score from the first stage is compared to a predicted base. The value-added measure was supplemented with measures of student attendance and dropout rates. Schools that showed positive growth were eligible for a reward. The prize pot was fixed which implied that the proportion of schools that could win was depended on the number of teacher and staff in the winning schools. Therefore, the relative ranking of schools was important too, and the program could be qualified as having tournament design features.

Methodology

To measure the impact of the program Ladd (1999) looks at the pass rate on the TAAS, she concentrates on the reading and math parts of the test. Although, she rather had used test scores there were no consistent set of test scores over time. A disadvantage of using the pass rate is that it does not measure improvement of student that already would pass without the program. Although as the pass rates are relatively low this might not be such a problem, she argues.

[16] The combination of primary and secondary schools is also called K-12 in the United States.

To address the counterfactual question she compared the results of the Dallas schools with five other large Texas Cities. In the regression estimation she tries to control for some of the differences between the cities. However, as the other five cities might also have employed programs the exact treatment effect stays unknown.

Results

The results point towards increased pass rates of 9 to 12 percent for reading and 12 to 17 percent for math in comparison to the other Texas cities for all years. Thus, relative to the base year of 1991 pass rate rose more in Dallas than in other Texas cities. The increased pass rates originated from the subgroups of Hispanics and whites. Strangely, black students did not show increased pass rates. There is however an important caveat. In the year before the program, which reflect gains before the incentive program, there is a similar positive effects found. Furthermore, the gains do not seem to increase over time, which is expected as students would have had more years of schooling under the program. Making a final conclusion of the precise effect that can be contributed to the incentive program difficult.

Commentary

Ladd (1999) seems to find a positive results but is bothered by several methodological issues that prevent a clear conclusion.

7.3 Randomized experiment

Randomized designs conduct experiments in which units are assigned randomly to the different conditions. This design has generally the best internal validity.

7.3.1. Dee and Keys (2004) - individual program

Description

Dee and Keys (2004) combine data from the Tennessee's Career Ladder Evaluation System with data from Project STAR. Because Project STAR insured that teachers and children were assigned randomly to classes it provided a great opportunity to get unbiased results. The paper includes 24.000 observations from children in Kindergarten through 4th grade[17].

The Career Ladder consisted of 5 steps. Each time the teachers wanted go up in the ladder, which was only possible at certain points in time, his performance was evaluated on

[17] This entails children from age 5 till 10.

six criteria. These criteria mostly covered teaching pedagogy and teacher skills. Evaluations were mostly based on class room observations by several evaluators. These evaluators were often teachers from other districts that reached the end of the ladder. The assessment also included information from the school principal, student questionnaires and peer questionnaires. Teachers also needed to pass a test. The ladder system provided incentives at step promotions between $1000 and $7000 dollar. There were no quotas on how many teachers could progress on the ladder.

Interesting design features included the following. One, it was completely based on subjective performance measures. Two, it combined several (subjective) data sources.

Methodology

Results were estimated by regression, as teachers and student were assigned randomly the results should be unbiased. Dee and Keys (2004) checked if the randomization of the STAR project was compromised, this was not the case.

Results

Students with career ladder teachers scored 3 percentile point higher on math tests and 2 percentile points higher on reading. However, only the math result was significant. This may be due the fact the project STAR was not designed to study the effect of the career ladder, it therefore had only weak statistical power to make meaningful interference. Interestingly, math teachers higher on the ladder did not significantly increase test score but teacher lower on the ladder did. In contrast, with reading teachers higher on the ladder did significantly increase test scores while teachers lower on the ladder did not.

Commentary

Dee and Keys (2004) find (positive) mixed result for a PRP program completely based on subjective performance measures. The results confirm firm the complexity of performance related pay programs, as it finds mixed results in the response of teachers.

7.3.2. Muralidharan and Sundararaman (2009) – individual and team incentive program

Description

Maralidharan et al. (2009) conducted an experiment for two years in the Andra Pradesh region located in India. They studied individual incentives, team incentives as well as the effect of increasing the input factors of a school. The experiments were held among rural primary schools, consisting of around 80 to 100 student and an average of three teachers. The average salary of teachers is around Rs. 8000 to 10.000 per month (per capita income per month in this region is around Rs 2000).

Teachers were awarded for improvement in students test scores. This was determined as followed. For the first year there was a baseline test held at the beginning of the school year (June/July) which was compared with an 'end of the year ' test (March/April). Furthermore, there was also a 5 percent improvement threshold in the first year. The authors were concerned that the results of the baseline test might have been artificially low due the fact that it was held after the summer vacation, with students tending to forget some of their knowledge in the summer vacation. The bonus payment could thus be described by the following formula:

$$Bonus = Rs.500 \cdot (\%\ gain\ in\ average\ test\ scores - 5\%)\ .$$

For the second year the 'end of the year' test of the first school year was used as the baseline test. Subsequently, the 5 percent threshold was removed.

The program design has some interesting features. First, the authors specifically choose to reward teachers on basis of the average test scores of students as they hoped to minimize possible 'threshold' effects. Second, the authors had complete control over the test, as it was specially designed for the experiment. As the authors had control over the content of the test they tried to control for 'teaching to the test' effects via an interesting way. The test contained both 'mechanical' and 'conceptual' questions. The mechanical questions were similar to text book exercises, whereby conceptual questions were unfamiliar. This approach seemed to work as conceptual questions, arguable sometimes easier, were more often answered incorrectly.[18] The idea of the authors was that when students started to make fewer mistakes both on answering the mechanical questions as well as on the conceptual questions it would be a sign of true learning, instead of some 'teaching to the test' effect. Three, precautions were taken against cheating, with external supervisors both present at test taking and when grading took place. Nonetheless, one class room and one entire school was caught cheating. Confirming the importance of monitoring schools with incentives carefully. Four,

[18] Example: Mechanical question: $\begin{array}{r} 34 \\ \underline{\times\ 5} \end{array}$, conceptual question: $8+8+8+8+8+8 = 8 \times \square$

In this example the mechanical question was answered correctly 43 percent of the time, while the conceptual question was answered correctly only 8 percent of the time.

to discourage teachers to exclude students with weak improvements, i.e. gaming of the test taking population, students that took the baseline test but were not present at the end of the year test were assigned a zero improvement score. For the second year a student who was present 'at the end of year test' in the first year but not present at the' end of the year test' in the second year was assigned a -5 percent rating. Thus, the cost of students dropping out was in both years the same. The authors capped negative performance of students also at the 5 percent level, so a teacher could never do worse than having a student drop out, giving little incentive for teachers to try to game the test taking population.

Methodology

For this experiment 500 representative schools located in the AP region were sampled. These groups were then randomly assigned to treatment groups for the different programs and a control group. The randomized design ensures that the estimated incentive effects, which are estimated by using a regression, are unbiased.

Results

The experiment provided the following results. In the individual program student on incentive schools scored 0.194 SD higher on math and 0.128 SD higher on reading after one year. The second year increased this to 0.239 SD higher for math and 0.152 SD higher for reading. The combined effect after two years of the program is that student scored 0.321 SD higher for math and 0.223 SD higher for reading. Note that this is not the sum of the separate year effects, as students forget some of the knowledge they have learned over time. For the school incentive program students in the first year scores 0.183 SD higher for math and 0.110 SD for reading. The second year effects are 0.128 SD higher for math and 0.042 SD higher for reading. Combining to a total effect over two years to 0.231 SD for math and 0.091 SD higher for reading. The positive results from both programs come from the entire distribution of students. Thus the individual incentive program is more effective, especially for reading.

Students in incentives schools (both individual as school) performed better at mechanical as well on the conceptual questions, letting the authors conclude that the increased results present actual learning instead of teaching to the test effects. Furthermore, on incentive schools students also scored higher on courses that were not part of the program. Scoring 0.11 to 0.18 SD higher than students in control groups. Pointing towards possible positive spillovers.

To gain insight in how teachers might have responded, class room observations as well as teachers interviews were undertaken. No significant changes were observed in class room observations. However, in the interviews teachers on incentive schools (both individual as school) responded significantly more often that they assigned more homework, conducted extra classes beyond regular school hours, gave more practice test and paid more attention to weaker children. The authors deemed these self reported measures credible as all of these activities show positive correlations, ranging between 0.079 and 0.183, with students test scores. The authors conclude that teachers changed their effectiveness in ways that could not be easily captured, even by observing the teacher. Lastly, it is found that both incentive programs were more cost effective than the input programs. The individual incentive program costs the same as the input program but test scores increased three times more compared with input schools. The group program was less costly but also had a smaller treatment effect than the individual program.

Commentary

Maralidharan et al. (2009) find a positive results for the performance related pay programs with a strong experimental design. One of the most interesting results is that there is also a positive spillover effect registered to non-incentive courses. Although the study may have great external validity for the AP region and India as a whole, it remains a question how much of these results can be generalized to a Western country. Sadly, the study did not further investigate the difference between the team and individual incentive program, besides the impact on test scores and cost effectiveness of the programs. Leaving interesting questions like: ‘Were teachers more favorable of the team or individual program?’, and ‘Was there a difference in reported cooperation levels between the two programs?’, unanswered.

7.3.3. Glewwe et al. (2010) – school based incentive program

Description

In 1998 a two year long incentive program was introduced in below average performing primary schools in (rural) Kenya. The typical school consisted of around 200 students and 12 teachers. Total teachers compensation, including other benefits besides wage, is around $2000 (in 1997). Which is five times annual GDP per capita. Teachers are paid on basis of education and experience and have strong union protection. Kenya suffered from high absent rates of teachers, with them being absent around 20 percent of the time.

The experiments provided rewards to teachers in grades 4 to 8 based on the performance of the school as a whole. Schools could either win in the category ‘top scoring school’ or ‘most improved’ but not in both. Improvement was measured by comparing test scores from district exams to a baseline year. This was done for all the seven subject which had district exams. Students also had to take other exams (KCPE and ICS exams) that were not tied to the incentive plan. They were used to control the effects of the study. In each category, three first prizes, three second prizes, three third prizes and three fourth prizes were award. At the end of each year there was an official award ceremony. Overall 24 out the 50 participating schools won prizes. The prizes ranged from 21% to 43% of teachers monthly salaries.

This incentive plan also has some interesting design features. First, it has two categories schools can compete in. Offering prospect to win an award for a wide range of schools. Relative good performing schools may have less room to improve student results than relative poor performing schools. On the other hand relative poor performing schools may have a smaller chance of winning in the top scoring category. Two, as seen before, schools were punished if student did not take the exam. Student who did not take the exam where assigned low (improvement) scores. Three, the program combines a tournament setup, as only a limited amount of schools can win a prize, with a school wide bonus. Lastly, the headmaster and three additional teachers of each incentive school were interviewed in the middle of the second year of the program.

Methodology

There was an initial group of 100 schools that had been selected for the program as the Ministry of Education considered these schools to be particularly in need of assistance. These schools then were randomly divided in a control and treatment group. The schools in the treatment schools participated in the teacher incentive program while the schools in the control group did not. The study has data for the pre-treatment year, the two treatment years and a post-treatment year. To estimating the treatment effects Glewwe et al. (2010) uses both a regression as well as a DID analysis. The randomization ensures both unbiased results from the regression as well as satisfy the assumption of common trends of DID analysis. For the analysis of the results student test scores were normalized. After normalization a normalized test score of 0.1 denoted that the student was 0.1 standard deviation above the mean score in the comparison schools. For a normal distribution this would mean moving from the 50th percentile to the 54th percentile.

Results

The following results were found. Test scores of district exams did significantly increase in participating schools. They found a positive treatment effect of 0.136 SD. The test scores on the other exams did also increase but not significantly. The gains however did not persist, after the incentive program ended the difference between students on incentive and control schools disappeared. Test scores during the incentive program especially increased in subjects that, arguable, involved much memorization, like geography, history and religion. Pointing strongly to short term' teaching to test' effects instead of actual learning. Test taking rates of the district exam also increased with 6 to 10 percent. Yet these effects also did not persist when the incentive program ended. Interestingly, anecdotal evidence reveals that not all teachers understood that having (bad) students drop out of the test reduced their chances of winning, as those students were assigned low scores. After the first year teachers seemed to realize this and test taking rates increased further in incentive schools. Furthermore, test taking on the other two test remained the same during the incentive program. Teachers behavior, teacher attendance, homework assignment and teaching methods also did not change much during the experiment, as observed by trained observers and information collected from students. Contrasting, in the second year interview 75 percent of the teachers claimed to have increased homework assignments as well as changed their attitude. A similar difference between official observations and self reported behavior of teacher was observed by Maralidharan et al. (2009).

Unsurprisingly, given the above results, test preparation sessions did increase. Which was also confirmed in the second year interview of teachers, as 88 percent of teachers reported that they increased preparation sessions. The interviews further revealed that teachers generally liked the program. Eighty-three percent thought the prize were awarded fairly in the first year of the program. Also 67 percent reported an increase in cooperation between teachers. Lastly, it is also worth mentioning that again one school was caught cheating.

Commentary

Glewwe et al. (2010) find far less favorable results as the positive results did not persist till after program. It seems that teachers focused on short term results instead of long term learning of students. This focus on the short term by teachers might have been caused by the fact that the program was initially announced as an one year program. Which was later extended to two years. Although other short term programs, for example Maralidharan et al. (2009) and Lavy (2002,2008), did not observe this behavior. Another explanation for the negative result is provided by Figlo and Kenny (2006). They argue that the negative results may be due the fact that all teachers received the same reward. However Lavy (2002), with a similar incentive program, did not find a negative result. Leaving this negative result somewhat unexplained. Maybe the most interesting finding of this paper is that teachers seemed to like the program as well as reporting increased cooperation levels. Finally, as the sample was pre-selected by the Ministry as well as the exotic location of the experiment external validity seems quite low.

8.4 Natural (quasi) field experiment

Natural field experiments study a naturally occurring differences between a treatment condition and a comparison condition. The researcher again as to deal with problems concerning the differences between treatment and control groups. However the researcher can study individuals without them knowing there are part of some special program, which may reduce possible Hawthorne effects. Hawthorne effects are used to describe short-lived increases in productivity due to change in behavior of individuals because of novelty or due the fact that they are being studied.

8.4.1. Martins (2010) – individual based program

Description

An interesting recent study is done by Martins (2010). In his paper he studies an education reform in Portugal that introduced incentive pay for teachers. The reform covered all public-sector schools, interestingly, two autonomous regions were not included. Those two regions did however introduce similar, lighter forms of the programs, making them an interesting comparison group. Private schools, were no new program was introduced, formed a second control group. Before the reform Portuguese teachers were paid via a single pay scale with automatic, tenure related progression. After the reform the pay scale was broken up into two scales. The first scale ends at €2000 per month, the second scale begins at €2500 per month. Teachers in the higher scale were supposed to play a special role in management and pedagogical tasks and gained the title ‘titular (head) teacher’. To move up in scales teachers were measured on several criteria. These criteria consisted of students performance (consisting of both internal (70%) as external grades (30%)), parents evaluations, attendance record (also at training sessions), management and pedagogical duties and involvement in research projects. Teachers who were in the higher scale needed to asses these criteria. Yet, teachers would only progress to the second scale if there was a vacancy open, as there are only a limited amount of upper scale teachers per school allowed. The amount of vacancies per school is determined every two years by the Ministry of Education. This setup introduced a tournament to the scheme as teacher will have to compete with each other for the limited amount of promotions to the upper scale.

The unique feature of this study is that it studies a reform and not a temporary pilot program like the other experiments. And unlike Atkinson et al. (2004) it has data on every high school and not just a subset. It therefore can draw conclusions for the entire population, i.e. average treatment effects, instead of drawing conclusion only for a certain subpopulation, i.e. an average treatment on the treated effects. Moreover, the incentive program affected all teachers and not only part of the teachers. For example, in most of the experimental studies only math and reading teachers were considered which leaves out other teachers like: history, or science teachers.

Methodology

To measure the impact of the education reform Martins (2010) used a difference-in-difference analysis. As mentioned above there are two control groups. The high schools in Azores and Madeira who implemented a lighter form of the reform and private schools who were not included in the reform. The data used in the study covers the period 2001-02 trough 2008-09, the reform was introduced in the school year of 2006-07. The DID analysis is appropriate as analysis of five years of data seem to indicate that the different groups are following common trends.

Results

The following results surfaced. Internal grades did not significantly change in comparison with the other two regions, Azores and Madeira. However (log) external grades significantly declined with 0,044 to 0,064. In comparison with the private schools internal as well as external grades decreased. The estimated were -0,020 to -0,029 and -0,052 to -0,070, respectively. In both cases evidence is found of internal grade inflation as the average gap between internal and external grades increases significantly. Interestingly, effect of grade inflation are found in 2008 and 2009 but not in 2007, the first year of the program.

Contributing to the evidence that teachers figure out over time how to play the system. Martins (2010) also provides evidence that teachers had reduced job satisfaction trough increased workloads and reduced cooperation. Subsequently, early retirements rates rose.

Commentary

Martins (2010) find, in stark contrast with most of the experimental studies, a negative results of the introduction of the PRP program, grades decreased and teachers had reduced satisfaction. This may point toward implementation issues for a permanent PRP program, unlike the temporarily nature of the experimental programs. First off all, as the reform changed the way teachers got paid there where contract negotiations between teacher unions and the government. The unions disliked the program because they thought it was unfair. Specifically, they disliked the fact that teachers with more experience and better academic qualifications might not be promoted to titular teacher as there was no vacancy available. While, at the same time, a less experience and worse qualified teacher might get promoted because his schools happens to have a vacancies available. The unions claimed that the program prevented 80 percent of the teachers to reach the position of titular teacher. At the end this dissatisfaction resulted in the largest demonstration involving a single sector since the introduction of democracy in Portugal in 1974. On 8 march 2008 100.000 teachers demonstrated against the plans, which was called the ‘March of Indignation’.[19] Thus before the program even started it had already a bad reputation, dooming it from the start. Second, a fundamental difference between the reform and the previous experiments is that the change in pay actually mattered for their ‘normal’ pay. In the experiments the performance-related pay part was always additional to the wage the teachers were already earning. Thus everything teachers might earn from the incentive program was a bonus. In the Portugal reform the performance-related pay program actually changed their ‘normal’ pay, as they no longer automatically progressed to the top of their pay. The Portuguese program also contrasted with the program in England studied by Atkinson et al. (2004). In England there was no quota and the upper pay scale was additional to the existing scale. In other words the program in England lacked the tournament structure. Hatry and Greiner (1984) argue that the use of quotas might have been a cause of the poor performance of old incentive plans in the United States. Third, the incentive program itself had a large amount of performance measures. The lack of focus might have made it hard for teachers to understand how to improve their

[19] http://www.eurofound.europa.eu/eiro/2008/04/articles/pt0804029i.htm accessed 17 august 2010.

performance. Furthermore, they might have felt that they have little control over their assessment. For example, teachers might have felt that they had little control over parents evaluations.

Concluding, the negative results of Martins (2010) should not be taken too lightly. However, over generalizing these findings towards the conclusion that PRP does not work is also not fair. As the design and implementation of the incentive program may not have been entirely optimal.

8.5 Summary and discussion

We have seen a wide range of studies ranging from individual to team based, from absolute to relative pay, from small to large bonuses, and from positive to negative results. Unfortunately, we come to the unsatisfying conclusion that their seems to be no clear recipe for success or failure. Five out of eleven studies found clear positive results, three out of eleven found partial positive results, and three out of eleven found clear negative results. The positive results are both under individual as team programs, and both under absolute as relative structures. Furthermore, the positive results are among different countries. These results thus cannot be easily dismissed. However, much of the same can be said of the negative results as they also are distributed among the range of different programs, and also among the different experimental designs. The studies of Elberts et al. (2002) and Martin (2010) may be especially interesting as part of their negative result may be explained by a sub-optimal design of the incentive programs. Thus providing valuable lessons on what not to do. Moreover, four studies find mixed results were a part of the teachers and students population did not respond to the program. To making things even more complicated there does not seem to be a pattern in which teachers and students do and do not respond to the PRP programs. For example, while Atkinson et al. (2004) find positive results under language teachers but no result under math teachers, Dee and Keys (2004) find the opposite results. All this makes drawing a general conclusion even more complicated. The same can be said about the absolute versus relative debate. The use of quotas in the Portugal program (Martin, 2010) seemed to have a negative effect on the program. However, other relative programs such as Ladd (1996) and Lavy (2002,2009) find positive results. Implying that relative payment may not be a problem per se. But maybe only relative pay in the form of quotas. Lastly, distorted behavior and cheating are very real concerns, although the magnitude in which they occur seem small. Three studies specifically report on teaching to test, only the study of Glewwe et al. (2010) in

Kenya finds evidence. Four studies report on cheating or grade inflation. One study (Lavy, 2008) does not find evidence of cheating, two studies only find one or two cases of cheating (Maralidharan et al. (2009) and Glewwe et al. (2010)). Only the study of Martins (2010) finds clear evidence of grade inflation. The result of Martins (2010) might be quite alarming as other studies might not have found results because of the short period in which they studied the program. As a final note I would like to mention three shortcomings. One, at least all experimental papers can only study the incentive effects of the program. They cannot study possible sorting and selection effects, as the programs only lasted for a short time. Lazear (2000) reported that these sorting and selection effects might be as large as the incentive effects. However, how this translates to performance-related pay for teachers remains to be seen. Should these effects be postive, as theory and private sector evidence indicates, than the estimated results of the experimental studies underestimate the positive effects of a possible (permanent) incentive program. Two, most studies only study part of the teachers population. For example only math and language teachers received rewards. As we previously mentioned there was already a diffference between how math and langauge teachers reacted. How other teachers thus are affected, like history, science, geoghrapy, or gym teachers remains an important, unanswered question. Three, the magnitude of programs and schools studied are too small to draw definite conclusions.

Chapter 8: Teachers support - Analysis SP survey

In this chapter I will try to do two things. One, I will try to uncover if Dutch teachers support the idea of PRP. Two, I will try to find out if teachers all think alike or differ in their opinion. For example, do males support the idea of PRP more than women? To my knowledge only Marsden (2000) has done a similar analysis.

To do this I have acquired a dataset from the Dutch Socialistic Party (SP)[20]. The dataset contains the results of a survey which was held under teachers in 2009. The survey contained a wide range of subjects, including ten questions about PRP. The SP has used this survey in their own publication 'de Leraar aan het woord' ('The Teacher Speaks'). '.. aan het woord' is a publication series in which trough the use of a survey a certain profession is asked for their (professional) opinion about the current state of their sector.[21] The survey was distributed by sending it to all primary and secondary school councils. An accompanying letter stressed the importance of filling out the survey and pointed out that previous surveys of the SP successfully changed public policy. After four weeks a reminder was send. The survey was also distributed to members of the SP who were known to work in education, it came with the request to make other teachers aware of the existence of the survey.

The results I received from the SP contained 1200 entries for secondary education and 1628 for primary education. These entries were already cleared of incomplete, double and other unusable entries. For the use of my analysis I further removed entries of special education, higher education, and entries of management that did not teach. At the end I had 967 entries left for secondary education and 1426 entries left for primary education. The entries for secondary education consisted of 219 vmbo school teachers, 208 havo and/or vwo[22] school teachers and 531 vmbo/havo/vwo school teachers.

This chapter will begin with a section about the representativeness of the survey, followed by the methodology and the results. Thereafter I will compare the answers UK teachers gave in Marsden (2000) with the answers in SP survey on three similar stated questions. This chapter will conclude with a discussion about the results.

[20] The SP is a left wing party which became increasingly popular in the last decade. Since their first election in parliament in 1994 the SP has been an opposition party.

[21] The publication series '… aan het woord' thus far include the professions of: police officer, general practitioner, youth worker, skipper, prison officer and teacher.

[22] for clarity, the havo/vwo schooltype consist of teachers that teach in a school that only offer havo or vwo and schools who offer both have and vwo.

8.1. Representativeness of the survey

When undertaking a survey it is important to have them filled out by a representative sample. If the sample is not representative the results may be biased, as the opinion of one group may not represent the opinion of another group. For example, men and women may differ in their opinion about PRP. Only asking men will clearly not represent women. In table 8.1 I compare the sample with the total population of teachers. For secondary school teachers males are overrepresented, the age distribution is however remarkable similar. If we look at the salary scales the sample also follows the general pattern of the total population, although the sample suffers from a substantial amount of missing values (10.9%). Lastly, teachers from the sample teach in smaller schools than the national average. For primary education there are no large differences between the sample and the total population while looking at gender and age. On salary scale there are again a lot of missing values (37.7%). In resemblance with the total population the majority of teachers are in the LA category. Finally, primary school teachers in the sample are from somewhat larger schools than the national average.

Although table 8.1 gives no sign for immediate trouble, we did not cover the most obvious concern yet. Namely, that the survey was mostly filled out by supporters of the SP. Supporter of the SP are more likely to be aware of the survey and more inclined to complete it. Given that the SP stands rather negative towards PRP[23] this may bias the results downwards. Obviously the dataset did not contain information about the political preferences of the participants, which made investigating this concern rather hard. The only thing that may give a slight clue is the geographical dispersion of the entries. For example, should 80% of the entries lay in the municipality of Oss (a cluster of SP voters) the participants are most likely biased. Thus to get some idea I did the following. For each entry (in the secondary education sample[24]) I looked up in which municipality the school was located, this was asked in the survey. I then for each entry added the percentage of SP voters in that particular municipality from the last national election in 2010. As you can see in table 9.1 the participating teachers came from schools who were located at municipalities that have on average (almost) as much SP voters as the national average. Obviously this proxy is far from

[23] See for example "CPB wil kille competitie tussen leraren" http://www.sp.nl/onderwijs/nieuwsberichten/7735/100601-cpb_wil_kille_competitie_tussen_leraren.html accessed at 8 September 2010 (in Dutch). Furthermore, it was together with the PVV the only party that did not include PRP in their election program. (CPB, 2010)

[24] Obviously this was a lot of tedious work, given that it is also a rather imperfect proxy I limited myself to the secondary education sample.

perfect as it does say nothing about the individual filling in the survey. Moreover, teachers may not work and live in the same municipality. But it is the best I could do.

Table 8.1.	Secondary education		Primary education	
	SP dataset (%)	Total population (in FTE) (%)	SP dataset (%)	Total population (in FTE) (%)
N	967	108.500 (pers.)	1426	146.800 (pers.)
Gender:				
male	63.5	56.2	24.6	23.0
female	35.8	43.8	75.4	77.0
Age:				
< 20	0.2	0.14	0.2	0.00
20 – 30	14.5	14.1	22.5	22.0
30 – 40	22.2	20.2	20.8	21.5
40 – 50	19.2	21.7	20.8	19.9
50 – 60	36.5	34.7	33.0	30.5
60 >	7.2	9.2	2.7	6.1
Salary Scale:				
LA	0.3	-	94.1	98.0
LB	61.3	60.6	5.3	2.0
LC	17.9	22.2	0.6	-
LD	20.5	16.9	-	-
LE	0.00	0.3	-	-
Other:				
Average amount of student per school	1068	1404	264	225
Election 2010, percentage SP voters	10.07	9.8	-	-

Note: N and gender taken from stamos.nl; includes teachers, principals and other personnel. Age and salary scale taken from 'Trends in beeld 2010' and WIO 2011, data consist of only teachers

To finish this section I looked a little bit further into the geographical dispersion of the entries. The entries are most likely clustered at certain schools, because a certain school may has an active SP member as teacher, has a more active council or teachers encourage each other more to complete the survey. This is something we indeed observed. An illustrative example is the municipality of Oosterhout (54.000 residents). Oosterhout has 22 entries in the secondary school sample, which is as much as Eindhoven (214.000 residents). Moreover, of the 22 entries 20 are from teachers that work in the same school type (i.e. havo/vwo).

Oosterhout has three high schools which of only one is a havo/vwo school. Which let us conclude that those 20 entries all come from the same high school. The results are thus clearly clustered at certain schools and thus may not be representative.

To conclude, the sample is by no means representative for the entire teacher population. However, gender, age and salary scales seem to be distributed approximately the same. Moreover, with 226 unique city/municipality entries their also seems to be a decent amount of geographical dispersion. The survey thus provides a great first step into discovering the opinion of Dutch teachers about PRP.

8.2 Methodology

To get a deeper insight into the answers of the teachers we estimate the following OLS regressions:

$$Y_i = \propto + \beta_1 \cdot School\ size + \beta_2 \cdot Urbanization + \beta_{3i} \cdot Gender_i + \beta_{4i} \cdot Age_i + \\ \beta_{5i} \cdot (Age_i \cdot Gender_i) + \beta_{6i} \cdot Schooltype_i{}^{25} + \varepsilon$$

With Y_i being the answers on the different question statements; school size indicating the amount of students in a school location; urbanization indicating the 'omgevingsaddressendichtheid' in the municipality were the school is located, this is the official definition used by Statistics Netherlands (CBS) (Dulk et al. 1992); gender is a dummy variable with male coded as one; age is also a dummy variable, the age category of 50-60 serves as the base; finally school type is a dummy variable for the three different school types, the vmbo/havo/vwo school type is the base category. Interaction terms between gender and age are also estimated, but are only shown if they produced significant results. If the results between the models with and without interaction terms substantially differed both are shown. I would like to warn the reader that these interaction terms might not be reliable as they have a limited amount of cases. In primary education the interaction terms were never significantly estimated, most likely due the limited amount of cases available.

Besides these regressions I also used two additional estimation specifications to check my results. First, I also estimated each regression per school type, the sample was thus split up into three parts. This estimation method was not used as main method because the sample size

[25] school type is only applicable for the secondary education sample. For secondary education it was also considered to include certification status into the regression. However a substantial part of the teachers had made use of the 'other' option, making it hard to include. Comparing just the means of the different groups did not indicate any large differences.

became too small for the interaction terms. Furthermore these models suffered from a substantial amount of cases with large standardized residuals, often resulting in insignificant models. If the main model estimated a significant school type dummy, the results of the separate regressions are reported. However caution is needed as these results might not be reliable. Two, I estimated the results via an ordinal regression. As the answers to the question statements are categorical this would be a more appropriate manner of estimation. However, ordinal regression is hard to interpret, therefore we use OLS as our main estimation method with the ordinal regression as robustness check. At last, the general coding of the answers is agree =1, neutral =2, and disagree = 3. The answer possibility no opinion is codes as missing. Furthermore in all of the tables the B's are the unstandardized coefficients and β's are the standardized coefficients.

8.3 Results

To give structure to the results I have divided the 10 items about PRP into 5 categories. Each category covers a certain aspect of PRP. Category one handles performance criteria in general, category two addresses subjective performance measures, category three looks at how teachers think about possible positive features of PRP, category four at a possible negative consequence of PRP, and category five concludes by looking at overall support for PRP. First the results for secondary education will be covered followed by the results for primary education.

8.3.1 Secondary education

Category 1: general performance criteria

Table 8.2.A: Category 1	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
42.G PRP causes teachers to be rewarded on basis of wrong criteria	41.3	19.8	37.4	1.3
42.A There are good criteria available on which teachers can be assessed if they are entitled to a higher pay scale	40.6	22.8	34.3	2

In this first category we make a further analysis of the question statements 42.G and 42.A. Both statements tell us something about how teachers think about (general) performance criteria. Note that the two questions have opposite phrasing. Thus at statement 42.G a lower score (i.e. agree) signals a more negative view. Specifically, those teachers think that a PRP program will use the wrong criteria to assess teachers. However, at statement 42.A a lower score (i.e. agree) signals a more positive view. Specifically, those teachers agree that there are good criteria available to assess a teachers performance for a PRP program.

If we look at the general results we see that each time a small majority of the teachers agreed with the two statements (table 8.2.A). Which is actually kind of surprising, as this means that teachers think that good criteria are available but they do not think these will be used in a PRP program. If we calculate the correlation coefficient between these two questions we find a correlation coefficient of 0.352. Indicating that indeed part of the teachers reasons this way. Which specific criteria are 'good' according to teachers remains an open question, and is left for future research.

The regression indicates that school size, gender and urbanization all do not have an impact (table 8.2.B). It does find an effect for the <20-30 age category. Specifically, teachers who are between <20-30 years old (relative to teachers who are 50-60 years old) seem to believe to a significant lesser extent that PRP will reward teachers on basis of wrong criteria. The regression also points to a significant interaction term for the first age category. This indicates that especially women drive the positive result of the 20-30 age category.

The vmbo variable also indicate a significant result. We therefore investigate the separate regressions. The effect found for <20-30 age category seems to be driven mostly by havo/vwo and vmbo/havo/vwo school teachers, explaining the significant result for the vmbo dummy in the main model. The results are also robust to the ordinal regression.

For the related question 42.A: “There are good criteria available on which teachers can be assessed if they are entitled to a higher pay scale” no significant model could be estimated. Thus teachers do not differ in their opinion that there are good criteria available.

Table 8.2.B: **Category 1**	Model 1: 42.G		Model 2: 42.A (non significant model)	
	B	β	B	β
Constant	1.694 (0.121)		1.801 (0.129)	
School size	0.000 (0.00)	-0.058	0.00 (0.00)	**0.080****
Urbanization	0.000 (0.00)	0.028	0.00 (0.00)	0.00
Gender	0.022 (0.097)	0.013	-0.098 (0.104)	-0.054
<20-30 versus 50-60 year	0.487 (0.13)	**0.207*****	-0.075 (0.140)	-0.030
30-40 versus 50-60 year	0.178 (0.12)	0.091	-0.170 (0.128)	-0.082
40-50 versus 50-60 year	0.075 (0.127)	0.036	-0.128 (0.137)	-0.059
60> versus 50-60 year	-0.398 (0.318)	-0.124	0.169 (0.341)	0.049
Interaction gender*age1	-0.402 (0.173)	**-0.126****	0.321 (0.186)	**0.095***
Interaction gender*age2	-0.092 (0.151)	-0.038	0.231 (0.161)	0.090
Interaction gender*age3	-0.003 (0.159)	-0.001	0.063 (0.170)	0.024
Interaction gender*age4	0.329 (0.34)	0.097	-0.018 (0.366)	-0.005
VMBO	0.166 (0.076)	**0.083****	0.118 (0.082)	0.057
HAVO/VWO	-0.087 (0.069)	-0.043	0.081 (0.074)	0.038

***p<0,01;**p<0.05; *p<0.1;model 1: $R^2 = 0.211$, model 2: $R^2 = 0.126$

Category 2: Subjective performance measures

Table 8.3.A: Category 2	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
42.B Principals are capable to assess teachers transparently	14.5	21.7	62.7	0.9
42.I PRP brings too much power to school principals	65.5	17.9	10	3.3

In this second category we again look at two question statements, namely 42.B and 42.I. These two statement will give us further insight in how teachers think about subjective performance measures. Again the two question are phrased differently. Thus, for question statement 42.B a lower score signal a positive view, i.e. the teacher thinks that a principal can assess him transparently. For question statement 42.I a lower score signals a negative view, i.e. principals will gain too much power by the introduction of a PRP program. It is striking that teachers answered both questions quite negatively (table 8.3.A).

For these two questions I used two models. The first model is the standard regression which is estimated each time. The second model includes, besides the standard items, a variable that indicates the relationship between teachers and management.

The first model (table 8.3.B) only shows a small negative results for the variable <20-30 vs. 50-60 year. Indicating that younger teachers are slightly more optimistic in the ability of principals to assess them. In the second model I added the variable: relationship management. In the survey teachers could rate their relationship with management on a 6 point likert-scale, from very poor (1) to very good (6). This variable shows a large significant results and removes the effects of the age categories. This is not really a surprising result, teachers who have a bad relationship with their management most likely have little faith in their ability to assess them.

Table: 8.4	
Age:	Rating:
20-30	4.67
30-40	4.45
40-50	4.15
50-60	4.17
60>	4.16

If we take a closer look at how teachers responded to this 'relationship' question, it is found that younger teachers are more positive about their relationship with the management than older teachers (table 9.4). Specifically, after the 30-40 age category teachers seem to be more negative. A t-test shows that the means between 20-30 (1% sig.) and 30-40 (5% sig.) age categories significantly differ with the other three age categories. Leading to the hypothesis that on average teachers apparently are getting somewhat disappointed in their management when the years

Table 8.3.B: **Category 2**	Model 1: 42.B		Model 2: 42.B + relationship		Model 1: 42.I		Model 2: 42.I + relationship	
	B	β	B	β	B	β	B	β
Constant	2.322 (0.096)		3.204 (0.124)		1.301 (0.089)		0.685 (0.116)	
School size	0.000 (0.00)	**0.081****	0.000 (0.00)	**0.058***	0.000 (0.00)	-0.057	0.000 (0.00)	-0.040
Urbanization	0.000 (0.00)	0.000	(0.000) (0.00)	0.002	0.000 (0.00)	0.024	(0.000) (0.00)	0.022
Gender	0.040 (0.051)	0.026	0.038 (0.049)	0.025	0.045 (0.047)	0.032	0.050 (0.046)	0.035
<20-30 versus 50-60 year	-0.140 (0.077)	**-0.066***	-0.045 (0.073)	-0.021	0.185 (0.071)	**0.096*****	0.119 (0.069)	**0.062***
30-40 versus 50-60 year	-0.092 (0.068)	-0.048	-0.038 (0.062)	-0.022	0.229 (0.060)	**0.143*****	0.195 (0.058)	**0.112*****
40-50 versus 50-60 year	0.089 (0.113)	0.021	0.089 (0.064)	0.067	0.186 (0.062)	**0.111*****	0.187 (0.060)	**0.111*****
60> versus 50-60 year	0.061 (0.102)	0.077	0.039 (0.096)	0.013	-0.005 (0.092)	-0.002	0.011 (0.089)	0.004
VMBO	0.136 (0.068)	**0.077****	0.124 (0.064)	**0.070***	0.018 (0.063)	0.011	0.029 (0.061)	0.018
HAVO/VWO	0.045 (0.062)	0.025	0.069 (0.059)	0.039	0.003 (0.057)	0.002	-0.009 (0.055)	-0.06
Relationship management			-.204 (0.19)	**-0.335*****			0.142 (0.18)	**0.258*****

***p<0.01;**p<0.05, *p<0.1; model 1: R^2 = 0.144; model 2: R^2 =0.360; model 3: R^2 = 0.167; model 4: R^2 =0.304

progress. If we stretch this to PRP this may indicate that subjective performance measures may more suitable to use with younger teachers, if used at all.

The results on the related question 42.I: “PRP brings too much power to school principals” are similar. Without taking the relationship variable into account, all age dummies are picked up by the regression, except for the 60> category. When taking the relationship variable into account this again reduces and a relative large effect is found for the relationship variable itself. The somewhat stronger results that are found for the age categories may be due the stronger question statement of question 42.I. All models are robust to the ordinal estimation method.

Category 3: positive features PRP

Table 8.5.A: Category 3	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
42.C PRP incentivize teachers to perform better	28.2	22.5	47.5	1.6
42.E PRP is more fair as additional effort and/or quality is rewarded	49	24.6	23.9	2.3
42.D With PRP teachers get a better career perspective	33.5	21.4	42.6	2.5

In category 3 we look at three question statements. All are phrased the same. At each question statement a lower score indicates a more positive view. Looking at the general results teachers do not think PRP will help their performance or their carrier perspective. However they do think it is more fair (table 8.5.A).

Again the regressions (8.5.B) find that younger teachers have a more positive view about PRP. At all three statements we also find an estimated effect for the first interaction term, indicating that once more women drive the effects that are found.

For the questions 42.E and 42.D there is also an effect estimated for the havo/vwo school type. The separate regressions of question 42.E seem to indicate that the results for young teachers are especially strong in havo/vwo schools. For question 42.D a strange result is found in the separate regressions. The significant have/vwo term seems to be mainly caused by a different effect in the older age and interaction term variables. In the separate regressions the 60> variable is found to be large, but insignificant, with a negative sign. Indicating that the old teachers think that PRP will increase their career perspective. At the same time a large (insignificant) positive interaction term is also estimated for this age category. Indicating that especially old women have this opinion. Strangely, for the other two school types reverse effect are found. Thus mostly old men are more positive about their career perspective, also insignificantly. As previous mentioned these models might not be reliable due a limited amount of observations. Other than that I do not have an explanation for these results. Moreover, only the vmbo/havo/vwo regression found an overall significant model. Ordinal regressions again shows similar results.

Table 8.5.B: Category 3	Model 1: 42C		Model 2: 42E		Model 3: 42D		Model 4: 42D + interaction	
	B	β	B	β	B	β	B	β
Constant	2.373 (0.086)		2.070 (0.090)		2.108 (0.069)		2.3135 (0.090)	
<20-30 versus 50-60 year	-0.517 (0.131)	**-0.213*** **	-0.620 (0.136)	**-0.245*** **	-0.322 (0.089)	**-0.129*** **	-0.498 (0.136)	**-0.199*** **
30-40 versus 50-60 year	-0.242 (0.122)	**-0.118** **	-0.088 (0.127)	-0.041	-0.208 (0.076)	**-0.099*** **	-0.147 (0.126)	-0.070
40-50 versus 50-60 year	0.221 (0.129)	**-0.102* **	-0.104 (0.136)	-0.046	-0.104 (0.080)	-0.047	-0.159 (0.134)	-0.072
60> versus 50-60 year	0.208 (0.329)	0.064	0.370 (0.343)	0.109	-0.022 (0.116)	-0.007	0.156 (0.338)	0.046
gender	-0.332 (0.175)	0.101	-0.017 (0.103)	-0.009	0.140 (0.060)	**0.076** **	0.100 (0.102)	0.055
Interaction gender*age1	0.075 (0.175)	**0.030* **	0.411 (0.182)	**0.120** **			0.335 (0.180)	**0.099** **
Interaction gender*age2	0.178 (0.161)	0.068	-0.102 (0.160)	-0.039			-0.111 (0.158)	-0.043
Interaction gender*age3	0.433 (0.191)	**0.158** **	0.001 (0.169)	0.000			0.084 (0.167)	0.031
Interaction gender*age4	-0.017 (0.350)	-0.005	-0.332 (0.365)	-0.093			-0.190 (0.360)	-0.053
VMBO	-0.102 (0.069)	-0.050	-0.116 (0.072)	-0.054	-0.105 (0.071)	-0.050	-0.105 (0.071)	-0.050
HAVO/VWO	0.133 (0.070)	0.064	0.155 (0.072)	**0.072** **	0.169 (0.072)	**0.079** **	0.170 (0.072)	**0.080** **
***p<0.01;**p<0.05; *p<0.; model 1: R^2 = 0.201, model 2: R^2 = 0.215, model 3: R^2 = 0.195 model 4: R^2 =0.211								

Category 4: negative consequence PRP

Table 8.6.A: category 4	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
42.H PRP causes division among teachers	68.9	18.7	10.3	1.8

A possible negative consequence of (individual) PRP is reduced cooperation. From question 42.H. it becomes clear that a large majority of teachers has this fear (table 8.6.A) The regression could not estimate a significant model, further indicating that this fear is widely accepted (table 8.6.B). The ordinal regression also did not estimate any significant results.

Table 8.6.B: category 4	Model 1: 43.H	
	B	β
Constant	1.363 (0.089)	
School size	0.000 (0.00)	**-0.064***
Urbanization	0.000 (0.00)	**0.065***
Gender	0.056 (0.047)	0.040
<20-30 versus 50-60 year	0.081 (0.070)	0.042
30-40 versus 50-60 year	0.066 (0.059)	0.042
40-50 versus 50-60 year	0.052 (0.062)	0.031
60> versus 50-60 year	-0.038 (0.092)	-0.015
VMBO	0.020 (0.063)	0.012
HAVO/VWO	-0.074 (0.057)	-0.046
*p<0.1; R^2 = 0.127		

Category 5: General opinion about PRP

Table: 8.7.A: category 5	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
41. Do you think that the salary scale should become depended on the individual performance of the teacher ?	38.7	18.8	32.9	9.7
42.F PRP can be used just as well in education as in other sectors.	33.5	16.5	48.1	1.7

In this last category we look at question 41:”Do you think that the salary scale should become depended on the individual performance of the teacher ?”. We also used a factor analysis on all the nine statement of question 42. The factor analysis extracted one factor, which we subsequently used as a dependent variable. A reliability analysis of this factor found a cronbach's alpha of 0,894. As our factor has a large number of items it may systematically raise the cronbach's alpha to large values (Cortina, 1993). Investigating the inter-item correlations shows that most correlations are well above the 'minimum' of 0.3. Although there are two instances when it falls below 0.3 (i.e. 0.281 and 0.266). The extracted factor may be interpreted as the overall opinion of teachers about PRP. A lower score indicates a more positive view. To finish this category we will also look at 42.F which asked if teachers think that PRP is suitable for the education sector or not.

Somewhat surprisingly, given the answers on the other statements, a small majority thinks that promotion in pay scale should become depended on performance. Strangely, on question 42.F a majority also answers that they do not think PRP is particularly well suited for the educational sector (table 8.7.A).

For question 41 the model indicates that there is no effect for age, gender, school size or urbanization. Also there is no difference between the teachers from different school types. In contrast with the other statements we also do not find a significant results for the younger age categories. However we do find a interaction effect of gender and age. Specifically men who fall in the 40-50 category answered the question more positive. Overall question 41 is a little bit of an anomaly. This may be caused by the differences in the question statements. While all the other statements specifically mention performance-related pay, question 41 does not. It could be that teachers do not link promotion in salary scale due individual performance explicitly to performance-related pay. Although question 41 actually states an example of performance-related pay.

When we use the factor as our dependent variable the regression finds a clear result for the lowest age category. However there is again a positive interaction term found. Which would indicate that young females offer more support for PRP, which is in line with previous results. Last, for question 42.F it is also found that the youngest age category is more positive. Indicating that younger teachers agree more with the statement that PRP fits the education sector just as well as any other sector.

All regressions indicate significant school type dummies. Therefore we investigate the separate regressions again. On each question similar results are found. The results for the youngest age and interaction terms seem to be driven mainly by havo/vwo schools. In vmbo schools the results is found that especially men in the 60> age category responded to the statements negatively. The vmbo/havo/vwo do not show any significant effects, or only in the lowest age category. The havo/vwo school type regressions are each time the only ones estimated to be significant.

Table: 8.7.A: **category 5**	Model 1: question 41		Model 2: 41 + interaction		Model 3: 41.F.		Model 4: factor analysis	
	B	β	B	β	B	β	B	β
Constant	2.059 (0.121)		2.119 (0.177)		2.111 (0.132)		0.136 (0.107)	
School size	0.000 (0.00)	-0.035	0.000 (0.00)	-0.037	0.000 (0.00)	0.032	0.00	**0.076****
Urbanization	0.000 (0.00)	0.045	0.000 (0.00)	0.040	0.000 (0.00)	-0.014	0.00	-0.034
Gender	-0.023 (0.065)	-0.013	-0.051 (0.109)	-0.028	0.102 (0.105)	0.055	-0.036 (0.123)	-0.017
<20-30 versus 50-60 year	-0.048 (0.096)	-0.019	0.178 (0.146)	0.071	-0.375 (0.140)	-**0.146*****	-0.588 (0.164)	**-0.205*****
30-40 versus 50-60 year	-0.174 (0.082)	**-0.082****	0.092 (0.133)	0.044	-0.013 (0.129)	-0.006	-0.270 (0.151)	**-0.114***
40-50 versus 50-60 year	-0.123 (0.086)	-0.055	-0.115 (0.142)	-0.052	0.008 (0.138)	0.004	-0.209 (0.161)	-0.084
60> versus 50-60 year	0.178 (0.130)	0.050	-0.404 (0.370)	-0.113	0.320 (0.347)	0.090	0.465 (0.386)	0.121
Interaction gender*age1			-0.281 (0.194)	-0.083	0.375 (0.188)	**0.107****	0.447 (0.219)	**0.114****
Interaction gender*age2			0.127 (0.168)	0.048	-0.100 (0.163)	-0.038	0.070 (0.189)	0.024
Interaction gender*age3			0.383 (0.178)	**0.140****	-0.177 (0.173)	-0.064	0.133 (0.200)	0.044
Interaction gender*age4			0.270 (0.396)	0.072	-0.251 (0.371)	-0.067	-0.318 (0.413)	-0.078
VMBO	0.098 (0.085)	0.047	-0.099 (0.085)	-0.047	-0.179 (0,083)	**-0.084****	-0.006 (0.095)	-0.003
HAVO/VWO	0.155 (0.078)	**0.072****	-0.156 (0.078)	**-0.073****	0.134 (0.076)	**0.061***	0.216 (0.087)	**0.089****

***p<0.01, **p<0.05, *p<0.01; model 1: R^2 = 0.135; model 2: R^2 = 0.175; model 3: R^2 = 0.196; model 4: 0.215

8.3.2 Primary education

The results for primary education are in general the same as for secondary education. The youngest age categories are estimated to have a significant effect. The specific results are reported below and commented on shortly.

Category 1: general performance criteria

Table 8.8.A: category 1	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
43.G PRP causes teachers to be rewarded on basis of wrong criteria	38.6	25.7	28.2	6.9
43.A There are good criteria available on which teachers can be assessed if they are entitled to a higher pay scale	38.5	18.6	39.0	3.2

Looking at the general results it is apparent that primary school teachers are a little more negative than their secondary counterparts. At question 43.G. there are a lot of teachers that voted neutral instead of disagree. Question 43.A. even shows a very small majority disagreeing with the statement (table 8.8.A).

The results of the regression seem similar to the results of secondary education. We see again no effect of school size and gender. But as before we see that young teachers are significantly more positive relative to the youngest age category at statement 43.G. Statement 43.A. again does not show a significant model (table 8.8.B). Both models are robust to the ordinal regression.

Category 2: Subjective performance measures

Table 8.9.A: category 2	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
43.B Principals are capable to assess teachers transparently	25.2	21.0	51.4	1.9
43.I PRP brings too much power to school principals	57.0	19.8	17.7	4.7

At category 2 we see an interesting development in the general results. It seems that primary teachers are more positive towards the involvement of their principals in a PPR. At question 43.B. 51.4 percent disagrees versus 62.7 percent for secondary teachers. For question 43.I. we

Table 8.8.B: **category 1**	Model 1: 42.G		Model 2: 42.A (non significant model)		Model 3: 42.A + interactie (non significant model)	
	B	β	B	β	B	β
Constant	1.738 (0.062)		1.939 (0.066)		1.902 (0.070)	
School size	0.000 (0.00)	-0.014	0.000 (0.00)	**0.072****	0.000 (0.00)	**0.073****
Urbanization	0.000 (0.00)	0.067*	0.000 (0.00)	-0.024	0.000 (0.00)	-0.025
Gender	0.048 (0.055)	0.024	-0.024 (0.058)	-0.012	-0.673 (0.308)	**0.324****
20-30 versus 50-60 year	0.221 (0.065)	**0.109*****	-0.023 (0.068)	-0.011	0.004 (0.077)	0.002
30-40 versus 50-60 year	0.207 (0.066)	**0.100*****	-0.080 (0.070)	-0.036	-0.045 (0.080)	-0.020
40-50 versus 50-60 year	-0.008 (0.065)	-0.004	-0.002 (0.069)	0.000	0.066 (0.081)	0.060
60> versus 20-50-60 year	0.231 (0.156)	0.0042	-0.040 (0.160)	-0.007	0.320 (0.224)	0.056
Interaction gender*age1					0.708 (0.340)	**0.144****
Interaction gender*age2					0.651 (0.337)	0.140*
Interaction gender*age3					0.507 (0.333)	0.121
Interaction gender*age4					0.750 (0.321)	**0.266****

***p<0.01; **p<0.05; model 1: R^2 = 0.180, model 3: R^2 = 0.106 model 3: R^2 = 0.076

observe a similar difference with 57 percent agreeing versus 65.5 percent. Although it is a clear difference the overall results still are clearly negative.

Table 8.9.B: category 2	Model 1: 42.B		Model 2: 42.B + relationship		Model 3: 42.I		Model 2: 42.I + relationship	
	B	β	B	β	B	β	B	β
Constant	2.279 (0.061)		3.580 (0.114)		1.553 (0.058)		0.555 (0.109)	
School size	0.000 (0.00)	0.028	0.000 (0.00)	-0.003	0.000 (0.00)	**-0.059****	0.000 (0.00)	-0.033
Urbanization	0.000 (0.00)	-0.017	0.000 (0.00)	-0.019	0.000 (0.00)	0.032	0.000 (0.00)	0.036
Gender	-0.082 (0.054)	-0.042	-0.056 (0.051)	-0.029	0.099 (0.062)	**0.054***	0.078 (0.049)	0.043
20-30 versus 50-60 year	-0.128 (0.063)	**-0.063****	-0.049 (0.060)	-0.024	0.142 (0.060)	**0.075****	0.081 (0.058)	0.043
30-40 versus 50-60 year	-0.065 (0.065)	-0.031	0.015 (0.062)	0.007	0.166 (0.061)	**0.087*****	0.098 (0.059)	**0.051***
40-50 versus 50-60 year	0.094 (0.064)	0.046	0.133 (0.060)	**0.064****	-0.029 (0.061)	-0.015	-0.059 (0.058)	-0.031
60> versus 50-60 year	-0.040 (0.146)	-0.008	-0.040 (0.139)	-0.008	-0.124 (0.141)	-0.025	-0.109 (0.135)	-0.022
Relationship Management			-0.275 (0.021)	**-0.346*****			0.210 (0.020)	**0.285*****
***p<0.01;*p<0.1; model 1: R^2 = 0.101, model 2: R^2 = 0.360, model 3: R^2 = 0.129, model 4: R^2 = 0.310								

The regressions again find the same results as before. Without taking relationship into account the younger age categories answered significantly more positive on the statement that principals can assess the teachers performance transparently. When we take the relationship variable into account, most of these effects disappear and the relationship variable shows large results, i.e. teachers who have scored their relationship with management more positive have a more positive view of their ability to assess them.

It is becoming rather predictive, but for statement 43.I. we also find similar results. The younger age categories disagree more with the statement that school principals may gain too much power. However, there is one difference, namely we observe also a gender effect. It seems that women are more afraid than men that principals gain too much power. However, the effect is limited. In the second model, the relationship variable takes over most of the explaining power.

If we again look a little bit closer to the relationship variable we get results that are also in line with previous results. The average score is 4.7 versus 4.3 for secondary education. The youngest age category score their relationship on average 4.86 while the oldest age category scores the relationship a 4.5. Primary school teachers thus seem to have a better relationship with their management. Over time the scores also decreases, but not as great as in secondary education. In contract with secondary education the t-test indicates that the hypothesis that the means between the age categories are the same cannot be rejected. Both models are robust to the ordinal regression.

Category 3: positive features PRP

Table 8.10.A: Category 3	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
43.C PRP incentivize teachers to perform better	33.2	18.5	42.3	2.1
43.D With PRP teachers get a better career perspective	45.1	17.4	33.9	3.1
43.E PRP is more fair as additional effort and/or quality is rewarded	48.8	15.6	32.4	2.5

Primary teachers are a little bit more positive about a possible incentive effect of PRP, with 33.2 percent agreeing versus 28.2 percent for the secondary education sample. Interestingly, primary school teachers agree a lot more with the statement that PRP will provide them a better carrier perspective, 45.1 agreeing versus 33.5 percent. Complementary with their secondary counterparts they also think performance-related pay is more fair (table 8.10.A).

By now it comes as no surprise that at all three statement the youngest age categories show significant results. Strangely, question C finds significant results for all age categories. Although the effects of the older age categories are a lot smaller (table 8.10.B). The ordinal regression does not shows this significant results for the 40-50 and 60> categories, but it does for the other two age categories. Indicating that it might have something to do with the OLS regression.

Table 8.10.B: Category 3	Model 1: 43.C		Model 2: 43.E		Model 2: 43.D	
	B	β	B	β	B	β
Constant	2.376 (0.063)		1.932 (0.066)		1.971 (0.065)	
School size	0.000 (0.00)	-0.004	0.00 (0.00)	0.029	0.000 (0.00)	0.022
Urbanization	0.000 (0.00)	-0.018	0.00 (0.00)	-0.029	0.000 (0.00)	-0.017
Gender	-0.019 (0.056)	-0.009	-0.002 (0.058)	0.000	0.047 (0.057)	0.023
20-30 versus 50-60 year	-0.475 (0.066)	**-0.222*****	0.253 (0.068)	**-0.118*****	-0.275 (0.068)	**-0.128*****
30-40 versus 50-60 year	-0.327 (0.068)	**-0.148*****	0.210 (0.069)	**-0.095*****	-0.267 (0.069)	**-0.121*****
40-50 versus 50-60 year	-0.180 (0.066)	**-0.083*****	-0.076 (0.068)	-0.035	0.026 (0.068)	0.012
60> versus 50-60 year	-0.371 (0.152)	**-0.067****	-0.282 (0.155)	-0.045	-0.126 (0.156)	-0.022

***p<0.01; **p<0.05; model 1: R^2 = 0.208, model 2: R^2 = 0.123, model 3: R^2 = 0.160

Category 4: negative consequence PRP

Table 8.11.A: Category 4	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
43.H PRP causes division among teachers	65.8	16.1	15.0	2.5

Category four has again no significant results (table 8.11.B) as no significant model can be estimated. Indicating again a widespread fear of division among teachers (table 8.11.A). Although the regression does finds a small gender effect indicating that men are little bit more afraid of division between teachers than women. There is also an effect estimated for 30-40 age category, them being less afraid of division between teachers (table 8.11.B). The ordinal regression does not estimate a gender effect, however it does also pick up the 30-40 age category.

Table 8.11.B: **Category 4**	Model 1: 43.F	
	B	β
Constant	1.415 (0.054)	
School size	0.00 (0.00)	-0.027
Urbanization	0.00 (0.00)	0.012
Gender	0.088 (0.048)	**0.051***
20-30 versus 50-60 year	0.071 (0.056)	0.040
30-40 versus 50-60 year	0.217 (0.057)	**-0.118*****
40-50 versus 50-60 year	0.015 (0.057)	0.008
60> versus 20-50-60 year	0.023 (0.127)	0.005

Category 5: General opinion about PRP

Table 8.12.A: Category 5	Agree (%)	Neutral (%)	Disagree (%)	No opinion (%)
42. Do you think that the salary scale should become depended on the individual performance of the teacher ?	38.8	-	40.0	20.7
43.F PRP can be used just as well in education as in other sectors.	41.9	14.3	41.2	2.0
Note that for question 42 the coding was agree= 1 and disagree = 2.				

For some unexplained reason there were no neutral entries at question 42. In contrast with the secondary education sample we observe a small majority disagreeing with idea of using PRP in there advancement of salary scales. Teachers are undecided on statement F. giving again a more positive signal than the secondary teachers (table 8.12.A). The reliability analysis for the extracted factor is 0.898, item-item correlations are all well above 0.3.

The regression for question 43.F did not estimate a significant model. All regressions find significant results for the first two age categories. Indicating that they are more positive towards PRP. In comparison with secondary education it is interesting to see that younger teacher also significantly answer more positive towards the question if their salary scale should become depended on their individual performance. Making the results in secondary education indeed an anomaly.

Table 8.12.B: **Category 5**	Model 1: 43.F (non sig model)		Model 2: 42		Model 3: factor	
	B	β	B	β	B	β
Constant	2.072 (0.067)		1.553 (0.041)		0.098 (0.132)	
School size	0.000 (0.00)	0.021	0.00 (0.00)	-0.048	0.000 (0.00)	0.073
Urbanization	0.000 (0.00)	-0.021	0.00 (0.00)	0.011	0.000 (0.00)	-0.059
Gender	0.007 (0.059)	0.003	-0.022 (0.035)	-0.019	-0.041 (0.107)	-0.018
20-30 versus 50-60 year	-0.146 (0.070)	**-0.066***	-0.156 (0.045)	**-0.131*****	-0.300 (0.134)	**-0.119****
30-40 versus 50-60 year	-0.191 (0.089)	**-0.084****	-0.137 (0.046)	**-0.114*****	-0.389 (0.129)	**-0.165*****
40-50 versus 50-60 year	-0.063 (0.082)	-0.028	-0.080 (0.042)	**-0.065***	0.003 (0.130)	0.001
60> versus 20- 50-60 year	-0.276 (0.158)	**-0.049***	0.007 (0.096)	-0.002	-0.234 (0.324)	-0.034
***p<0.01;**p<0.05 *p<0.1; model 1: R^2 = 0.092, model 2: R^2 = 0.134 model 3: R^2 = 0.194,						

8.4 Comparison with Marsden (2000)

A few questions in Marsden (2000) may be compared to questions asked in the SP survey. This may give some context to the results. However, as different nuances in the questions may invoke very different responses, not too much weight should be given to this comparison.

Question 21 in the survey of Marsden (2000) asked: "linking pay with the performance review will result in a fairer allocation of pay". In contrast with Dutch teachers, UK teachers were quite negative. Seventy percent of the UK teachers disagreed with this statement, 15 percent was neutral and 11 percent agreed. In comparison we have seen that around 50 percent of the Dutch teachers agreed with the similar statement 42(3) E.

The survey of Marsden (2000) also contained the question: "PRP will cause jealousies". We may compare this to our statement: "PRP will cause division among teachers". In Marsden around 90 percent of the teachers agreed that PRP will cause jealousies. This is even more negative than Dutch teachers, were between 65-70 percent thought PRP will cause division among teachers.

Another question in Marsden (2000) was: "It is very hard for teachers like me to improve our performance because we already work as hard as we possibly can". Which we may compare to: "PRP will cause teachers to perform better". Eighty-eight percent of the UK teachers agreed with the statement. While around 45 percent of the Dutch teachers agreed that they will not perform better under PRP, they are again less negative.

This quick comparison indicates that the UK teachers are possibly even more skeptical of performance-related pay than Dutch teachers. It is therefore hopeful to note that despite this lack of support a (mixed) positive effect was found for the PRP program that was introduced in the UK (section 7.2.3). Indicating that a PRP program may possibly overcome initial skepticism of teachers.

8.5 Discussion

The discussion about the results can be rather brief. The general results seem to indicate that teachers are somewhat divided in their support for PRP, but are certainly not so unanimous against PRP as sometimes is left to believe. The deepest concern of teachers seems to be that a PRP program will cause division among teachers. A more in-depth analysis unveiled that younger teachers are more supportive of PRP than older teachers. The regression found across almost all statements and both in primary as secondary education significant results for the 20-30 age category and in lesser extent for the 30-40 age category. For readers who might enjoy more specific results than the outcomes of the regression analysis, in Appendix C I have included tables that directly compare the answers between teacher from the 20-30 and 50-60 age categories. While it confirms again our main results, it shows that the differences between the two age categories can rise up to 20 percent agreeing (disagreeing) more with certain statements. I deem this very substantial as it causes going from a majority disagreeing (agreeing) with a statement to a majority agreeing (disagreeing) with a statement.

For policymakers this might be a welcoming message. We have seen in section 2.3 that in the coming years the teacher population will rejuvenate. The share of young teachers will increase while the share of old teacher decreases. According to the results that are found this should be positive for the support for PRP among teachers. The survey did not contain information on precisely why young teachers are more supportive of PRP. Interestingly, Maralidharan et al. (2009) found in their experiment that young teachers indeed responded better to a PRP program than older teachers. They thought this might be due younger teachers responding better to new policies than older teachers. Although it is the only study that specifically reported such an effect. The results are also partly in line with Marsden (2000). Marsden (2000) also found that younger teachers were more convinced that the new performance-related pay program would offer them a better carrier perspective. However, in contrast with our results, on other statements no differences was found between younger and older teachers.

Urbanization, school size and gender do not seem to have any impact. Especially the absence of an effect of school size is notable. You might expect that in larger schools teachers might be less concerned about division between teachers or think more negatively about the ability of principals to assess them. Yet no effects are found. Which might actually be encouraging for policymakers. Should the support for PRP differ between school sizes, urban location or gender it could severely complicate a national policy. Suppose for example that teachers in small schools would be very positive towards subjective measures while teachers in large schools would be very negative. Policy makers then would have to differentiate between large and small schools when making their public policy. Not to speak of a possible gender effect. Although, as we may learn from the UK example, ex-ante support and ex-post support / behavior obviously may differ.

Another notable result was that teachers seem very opposed towards subjective performance measures, although in line with the general results younger teachers are a bit more positive. Dutch policy makers thus are probably better off with using objective measures.

The result that young women's sometimes drove the positive effects we found for the 20-30 age category is also noteworthy. Recent literature seem to indicate that men perform better in competition than women. This literature also indicates that when given the choose men are more likely to choose a competitive environment than women (Gneezy and Croson, 2009). However much is yet unknown, and the available research is too small to draw general conclusions. Lavy (2009) provides a great example by not finding a difference in performance between the genders in an individual teacher incentive program (see also section 7.2.2.) Moreover, there is even less research available on how team competition influences a possible gender effect.[26]

Finally for secondary education we cannot make any strong statements about the effect of the different school types. Relative small effects are sometimes picked up, but they never get very significant (1%). Separate regressions seem to indicate that especially havo/vwo schools drive the result we found for the 20-30 age category. However, caution is advised as these models might not be reliable. Future research could investigate if there is indeed a difference between the reaction of the different school types towards PRP.

[26] the few studies that do exist are to my knowledge: Delfgaauw et al. (2009), Stenzel and Kübler (2005) and Dargnies (2009).

Chapter 9: Summary, Conclusion and Policy Recommendations

In the introduction of this thesis I noted that the new Dutch government will, starting in 2015, structurally invest €250 million in performance-related pay for teachers. Throughout this thesis the readers has been introduced to the world of performance-related pay for teachers. The general theory of performance-related pay has been discussed as well as specific problems that may arise due the specific nature of the teaching profession. Furthermore, this thesis has looked into the current empirical literature and investigated the support among Dutch teachers for a performance-related pay program. In this last chapter the main findings are summarized. I will conclude with some policy recommendations and answer the main question of this thesis: Should performance-related pay for teacher be introduced in the Netherlands?

9.1 Summary

In chapters 2 and 3 we have looked at why the current pay system may need to be changed. In chapter 2 we described the current and future state of Dutch labor market for teachers. While in the past the shortage of teachers was considered to be a major problem recent estimations seem to indicate that it might not be that bad. The age distribution of teachers will change the coming years, many old teachers will be replaced by younger ones. Teacher salary is not particular bad, but also not particular good. Teachers themselves however are quite unsatisfied about their salary. In chapter 3 we saw that teachers were highly important to student results. Good teachers are not identified by experience, academic achievements or certification. Precisely the criteria on which the current system pays. The current system therefore does not help to retain or attract top quality teachers.

In chapter 4, we described why performance-related pay may change this by looking at the literature of performance-related pay in general. The theory describes that performance-related pay will have two advantage to the current system. First, it will create incentives. It therefore will encourage teachers to put in more effort. Second, it will have sorting and selection effects. High quality teachers will be attracted to the profession while low quality teachers are discouraged. However, performance-related pay comes with his own disadvantages. For instances the effort of teachers is hard to measure, therefore limiting the role incentives can play. Also, teachers may engage in multiple tasks. All tasks should be rewarded equally or else teachers may neglect some of his tasks. For example, the teacher may only concentrate on test preparations, i.e. 'teaching to the test'. Other distorted behavior like downright cheating is also a possibility. An incentive program therefore always has to be monitored and safeguarded against opportunistic behavior. Another lesson is that including multiple information sources may make the contract more efficient. Lastly, we discussed the possibilities of team and relative pay. Overall we can conclude that while introducing performance related pay might be a good idea, its scope should be limited when used in the teaching profession.

Chapter 5 has examined the interactions between performance-related pay and psychological factors by examining the psychological as well as the economic literature. We concluded that the presumed negative interaction between intrinsic motivation and monetary incentives might not exist. Which does not mean that other factors may not play a role. Other psychological factors such as reciprocity, social norms, status and fairness all may interact with monetary incentives via one way or another. Policymakers therefore should always be on the lookout for possible unwanted interactions.

Chapter 6 investigated the literature on how teachers performance might be measured. Objective measures include test scores and pass rates and have the disadvantage that they may induce teachers to concentrate too much on test results of students. When using objective measures there should be controls for social-economic factors, for example via value-added models. However, this may make measuring teachers performance rather complicated. Moreover, it seems that value-added models can only reliable identify the top and the bottom of the teacher distribution. Subjective measures can complement the objective measures, although they also have it shortcomings. Subjective measures may induce non-productive influence activities by teachers as well as induce favoritism by principals. Subjective measures also only seem to be able to differentiate between the top and worse performing teachers. Therefore making relative pay more suitable to use with current available criteria. Both criteria are imperfect, making measuring the performance of teachers the greatest challenge for introducing an incentive program for teachers.

In chapter 7 eleven papers are discussed that studied previous implemented performance-related pay programs. However, they do not offer much help in identifying a successful strategy when designing a performance-related pay program. The majority of papers finds a positive effect or mixed positive effects, although there are also so papers that find negative effects. The literature provides little help in whether to choose relative or absolute pay or individual versus team pay. All have caused positive effects but all also have caused negative effects.

In Chapter 8 we made use of a survey from the SP to investigate how much and which teachers support the idea of performance-related pay. Teachers seem to be rather divided between their support for performance-related pay. Teachers seem most concerned about the transparency and power that principals may receive in a performance-related pay program. Moreover, there seems to be a wide spread fear among teachers that an incentive program will cause division among teachers. Furthermore, a majority of secondary school teachers does not believe that they will perform better or that they get a better carrier perspective. On a positive note, teachers do think that an incentive program will be more fair. They also think that good criteria are available to determine their performance. However a majority thinks that performance-related pay program will not use these criteria's. There is not much difference in the answers between primary and secondary school teachers. The most notable difference is that primary school teachers do think that an incentive program will give them a better carrier perspective.

An in-depth analysis indicated that school size, urbanization and gender do not have any effect on teachers support for performance-related pay. However, we did find that younger teachers, specifically teachers between <20-40 years old, are more positive toward the ideas of performance-related pay than older teachers. With the inflow of young and outflow of old teachers in the near future this might be a welcoming message for performance-related pay. These results were both found in primary as in secondary education. A noteworthy finding was that in some of the statements the positive effect of the youngest age category was mainly driven by women. Especially in light of recent literature about gender differences in a competitive environment.

9.2 Conclusion and Policy Recommendations

So should the government go through with their plans to invest in performance-related pay for teachers? The answer to that question is a cliché. Namely, yes if properly designed. Now, ask any random economist if a performance-related program of *any* kind will work and he will probably give the same answer. But this thesis hopefully made clear that the unique situation of teachers does not necessarily prevent a successful incentive program. Which brings us to the real question: what is the 'proper' design for a teacher incentive program? This is a much harder question to answer as the literature of teacher incentives does not has a clear answer ready yet. The precise workings are not yet clear and there is no recipe for success available. With most of the problems originating from the fact that it is hard to measure teachers performance. Teachers may be skeptical but are not unanimously against PRP. Moreover, the results of my analysis indicates that support for PRP may increase substantially the coming years as many young teachers will enter the profession. Finally, I would like to give at least a few policy recommendations that may help policymakers.

One, start-up pilot programs. One thing we learned from the empirical evidence is that the response of teachers to the incentive programs were precarious. For example, sometimes math teachers reacted and language teachers did not and vice versa. It seems that specific circumstances might influence the working of performance-related pay programs. Thus before initiating some kind of national policy it might be wise to gain information on how teachers in the Netherlands specifically react to incentive programs. It could be that specific characteristics in the Dutch education system influence the working of a performance-related pay program. Moreover, introducing a successful program most likely will come with some trial and error. Doing this in a pilot program might be better than in a national policy. Furthermore, it introduces teachers to performance-related pay and, if successful, may make it easier to implement a national policy.

Two, introduce the performance-related part as *additional* variable part in the salary of the teacher. One, because the output of teachers is hard to measure the theory prescribes that the variable part cannot play a very large role anyway. Two, in the SP survey a small majority of teachers in secondary education, and almost an equal amount of teachers in primary education, agreed that their salary scale should become more depended on individual performance. Introducing performance-related pay as *additional* is, in my opinion, a more favorable way for teachers to have their individual performance count in their pay. Teachers would still automatically progress in the salary scale, however they would now also have the possibility to earn something extra. It is my hypothesis that this way of introducing performance related pay would have stronger support among teachers. Moreover, it may gives a nuance to the public debate of performance-related pay for teachers. Opponents try to give the impression that performance related pay will change teachers salary into some variable salary, like a fruit picker[27]. However, like we saw in the empirical literature and as I propose,

[27] see for example http://www.fd.nl/artikel/20683205/onderwijs-gaat-goede-leraren-belonen-maar-weet-niet (in Dutch) or http://www.trouw.nl/krantenarchief/2010/04/16/3043207/De_keerzijde_van_een_bonus_voor_docenten.html

teachers kept earning the same base wage as before, only good performing teacher were rewarded, at the end of the year, with a bonus. This would of course mean an actual investment in teachers and their salary, and not just a rearranging of funds. As a downside this way of introducing performance-related pay may have a limited amount of possible sorting and selection effects. In Martins (2010) were base pay did became intertwined with variable pay a negative result was found. Conforming that it might be better to keep those two apart. A quick back-of-the-envelope calculation indicates that this bonus for successful teachers might lay on average around €1200[28].

As a side note, this does not necessarily mean that the current base pay should not be changed. Although beyond the scope of this thesis, there is part of the literature that also argues that current *base* pay is in need of change. They propose a knowledge-and skills based pay plan for teachers *base* pay (Odden, 2008).

Three, in measuring teachers performance the focus should be on objective performance measures. Teachers are highly skeptical of subjective measures, and do not seem to support it.

Four, relative pay seems currently more suitable to use in a teacher incentive program than absolute pay, as the performance criteria can only reliable assess the top and the bottom of performing teachers. However, in the empirical literature we have seen that also absolute programs have achieved positive results. Leaving the option between the two still open.

Five, the debate between individual and team pay is still undecided. Both seem able to work (and fail). Individual incentives do seem to find a greater impact on scholastic results of students. The decision between individual and team pay may also be depended on the support of teachers, administrators and unions. Sadly, the SP survey could not give further insight into this.

Six, policymakers should beforehand decide what they would like to achieve besides offering teachers the opportunity to earn more wage. Should the program focus on poor performing students, then the use of pass rates might be useful. Or should the program focus on all students, then some average of test scores might be useful. Of course one could also make use of a very easy criteria. For example, reduce the amount of drop-outs. However, do not expect than that scholastic results of the students will improve.

[28] Government indicates to invest €250 million, the total employment in Dutch education is around 420.000 jobs (includes all personnel and all educational sectors, stamos.nl). Assuming that halve of this personnel gets rewarded, bonus per person would be: $\frac{250.000.000}{210.000} = 1190.48$.

Seven, policymakers should be aware of some practical issues. If an incentive program will make use of test scores that control for social-economic factors information systems of schools need to be of high level. For example, can current information systems link specific students to teachers over multiple years? The use of test scores may further implicate a greater use of standardized tests. These test need to be developed, may have implementation issues etc. Which brings me back to the first point, you might need pilot programs to smooth things out.

To conclude it all, the 'proper' design of a teacher incentive program is not simple. I therefore would not recommend the introduction of national performance related pay program for teachers any time soon. Luckily, the government will already wait till 2015 before it start to invest. This will give economist and policymakers time to prepare a 'proper' design and possibly make use of the growing literature about teacher incentives. If this is accomplished by 2015 remains to be seen. However, politicians should only introduce a national wide performance-related pay program if they are convinced it has a proper design. Otherwise it will waste the potential performance-related pay has and most likely a lot of money with it.

When the financial crisis hit economists were blamed for not seeing it coming. Turning to economics to predict the future is foolish, turning to anyone to predict the future is foolish for that matter. Or as a old Arab proverb goes: "He who foretells the future lies, even if he tells the truth". While predicting the future can be removed from the job description, advising on a performance-related pay program for teachers isn't. That a properly designed incentive system can increase productivity might be one of the main claims of today's economists, and maybe also the most contested one. We have seen in this thesis that designing an incentive program for teachers will be a difficult challenge. However, unlike predicting the future, this is actually a challenge that economists might be able to meet.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics , 25* (1), 95-134.

Atkinson, A., Burgess, S., & Croxson, B. (2004). Evaluating the impact of performance-related pay for Teachers in England. *CMPO Working Paper 04/113* .

Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review , 20*, 51-61.

Barlevy, G., & Derek, N. (2009). Pay of Percentile. *IZA discussion paper* (No. 4383).

Burgess, S., Croxson, B., Gregg, P., & Propper, C. (2001). The intricasies of the Relationship Between Pay and Performance for Teachers: Do Teachers respond to performance related pay schemes. *CMPO Working Paper series* (01/35).

Cortina, J. (1993). What is coefficient alpha? An examination of of theory and applications. *Journal of Applied Psychology* , 98-104.

CPB. (2010). *Bijzondere Publicatie No. 85: Keuzes in kaart 2011-2015: Effecten van negen verkiezinsprogramma's op economie en milieu.*

Dargnies, M. (2009). Does Team Competition eliminate the gender gap in entry in competitive environments. *Working Paper University Paris* .

de Commissie Leraren. (2007). *LeerKracht!* Ministirie van Onderwijs, Cultuur en Wetenschappen.

Deci, E. L., Ryan, R. M., & Koester, R. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin , 125* (6), 627-668.

Deci, E. (1971). The Effects of Externally Mediated Rewards on Intrinsic Motivation. *Journal of Personality and Social Psychology* (18), 105-115.

Dee, T. S., & Keys, B. J. (2004). Merit Pay reward good teachers? Evidence from a Randomized experiment. *Journal of Public Policy Analysis and Management , 23* (3), 471-488.

Delfgaauw, J., Dur, R., Sol, J., & Verbeke, W. (2009). Tournament incentives in the field: Gender differences in the workplace. *Working Papers Tinbergen Institute* .

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. *CEPR Discussion Paper No. 6059* .

Dulk, d. C., Stadt, v. d., & Vliegen, J. (1992). Een nieuwe maatstaf voor stedelijkheid: de omgevingsadressendichtheid. *Maandstatestiek bevolking* (7).

Ecorys, CentER data, QQQ Delft. (2006). *De toekomstige arbeidsmarkt voor onderwijspersoneel tot 2015.* Ministirie Ondewijs, Cultuur en Wetenschap.

Eisenberger, R., & Cameron, J. (1996). Detrimental Effects of Reward. *American Psychologist , 51* (11), 1153-1166.

Eisenberger, R., Rhoades, L., & Cameron, J. (1999). Does Pay for Performance Increase or Decrease Perceived Self-Determination and Intrinsic Motivation. *Journal of Personality and Social Psychology , 77* (5), 1026-1040.

Elberts, R., Hollenbeck, K., & Stone, J. (2002). Teacher performance incentives and student outcomes. *The Journal of Human Resources , 37* (4), 913-927.

Fehr, E., & Falk, A. (2001). Psychological Foundations of Incentives. *Working Paper No.95 Univeristy of Zurich* .

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives , 14* (3), 159-181.

Figlio, D. N., & Getzler, L. S. (2002). Accountability, Ability and Disability: Gaming the System. *NBER Working Paper Series* (9307).

Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics , 91*, 901-914.

Gibbons, R. (1998). Incentives in Organizations. *Journal of Economic Perspectives , 12* (4), 115-132.

Glewwe, P., Illias, N., & Kremer, M. (2010). Teacher Incentives. *American Economic Journal: Applied Economics , 2* (3), 205-227.

Gneezy, U., & Croson, R. (2009). Gender Differences in Preferences. *Journal of Economic Litarature , 47* (2), 448-474.

Green, J. R., & Stokey, N. L. (1983). A Comparison of Tournaments and Contracts. *Journal of Political Economy , 91* (3).

Hanushek, E., & Rivkin, S. (2006). *Handbook of the Economics of Education, Volume 2, Chapter 18.*

Hatry, H. P., & Greiner, J. M. (1984). *Issues in Teacher incentive plans.* National Inst. of Education.

Jacob, B. A. (2005). Accountability, incentive and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economy , 89*, 761-796.

Jacob, B. A., & Lefgren, L. (2005). Principals as Agents: Subjective Performance Measurement in Education. *KSG Faculty Research Working Paper Series* (RWP05-040).

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics* , 843-877.

Jenkins, G., Gupta, N., Atul, M., & Shaw, J. (1998). Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research. *Journal of Applied Psychology , 83* (5), 777-787.

Johnson, S. M. (1986). Incentives for Teachers: What Motivates What Matters. *Educational Administration Quarterly , 22* (3), 54-79.

Kane, T. J., & Staiger, D. (2002). Volatility in School Test Scores: Implications for Test-Based Accountability Systems. *Brookings Papers of Education Policy* , 235-283.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review , 27*, 615-631.

Kunz, A. H., & Pfaff, D. (2002). Agency theory, performance evaluation, and the hypothetical construct of intrinsic motivation. *Accounting, Organizations and Society* (27), 275-295.

Ladd, H. F. (1999). The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes. *Economics of Education Review , 18*, 1-16.

Ladd, H. F., & Clotfelter, C. (1996). *Holding Schools Accountable: Performance-Based reform in Education.* The Brookings Insitution.

Lavy, V. (2002). Evaluating the effect of Teachers Group Performance Incentives on Pupil Achievement. *Journal of Political Economy , 110* (6).

Lavy, V. (2008). Gender Differences in market competitiveness in a real workplace: Evidence from Performance-Based pay tournaments among teachers. *NBER Woring Paper 14338* .

Lavy, V. (2009). Performance Pay and Teachers Effort, Productivity and Grading Ethics. *The American Economic Review , 99* (5), 1979-2021.

Lazear, E. P. (2000). Performance Pay and Productivity. *The American Economic Review , 90* (5), 1346-1361.

Lazear, E. P. (1986). Salaries and Piece Rates. *Journal of Business , 59* (3), 405-431.

Lazear, E. P. (2003). Teacher incentives. *Swedisch Economic Policy Review , 10*, 179-214.

Lazear, E. P., & Rosen, S. (1981). Rank-order Tournaments as Optimum Labor Contracts. *Journal of Political Economy , 89* (5).

Marsden, D. (2000). *Teachers Before the 'Threshold'.* Centre for Economic Performance (Londen School of Economics.

Martins, P. S. (2010). Teacher Incentives, Student Achievement and Grade Inflation. *IZA Discussion Paper No. 4051* .

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating Value-Added models for teaching accountability.* Rand Cooperation.

Milanowski, A. (2008). How to pay teachers for student performance outcomes. *Series of Teacher compensation papers from the Univeristy of Wisconsin CPRE Group* .

Milgrom, P. R., & Roberts, J. (1992). *Economics, organization and management.* Pearson Education.

Ministerie van Onderwijs, C. e. (2009). *Werken in het onderwijs 2010.*

Ministerie van Onderwijs, C. e. (2010). *Werken in het onderwijs 2011.*

Muralidharan, K., & Sundararaman, V. (2009). Teacher Performance Pay: Experimental Evidence from India. *NBER Working Paper 15323* .

Murnane, R., & Cohen, D. (1986). Merit Pay and the Evaluation problem: Why most merit pay plans fail and few survive. *Harvard Educational Review , 56* (1).

Neal, D. S. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *The Review of Economics and Statistics , 92* (2), 263-283.

Odden, A. (2008). New Teachers pay structures: The compensation side of the Strategic Management of Human Capital. *CPRE* .

Podgursky, M. J., & Springer, M. G. (2006). Teachers Performance Pay: A Review. *Working Paper: www.performanceincetives.org* .

Predergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature , 37* (1), 7-63.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teacher, Schools, and Academic Achievement. *Econometrica , 73* (2), 417-458.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized studies. *Journal of Educational Psychology , 66* (5), 688-701.

Ryan, R. M., Koestner, R., & Deci, E. L. (1991). Ego-involved persistence: when free-choice behavior is not intrinsically motivated. *Motivation and Emotion* (15), 185-205.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

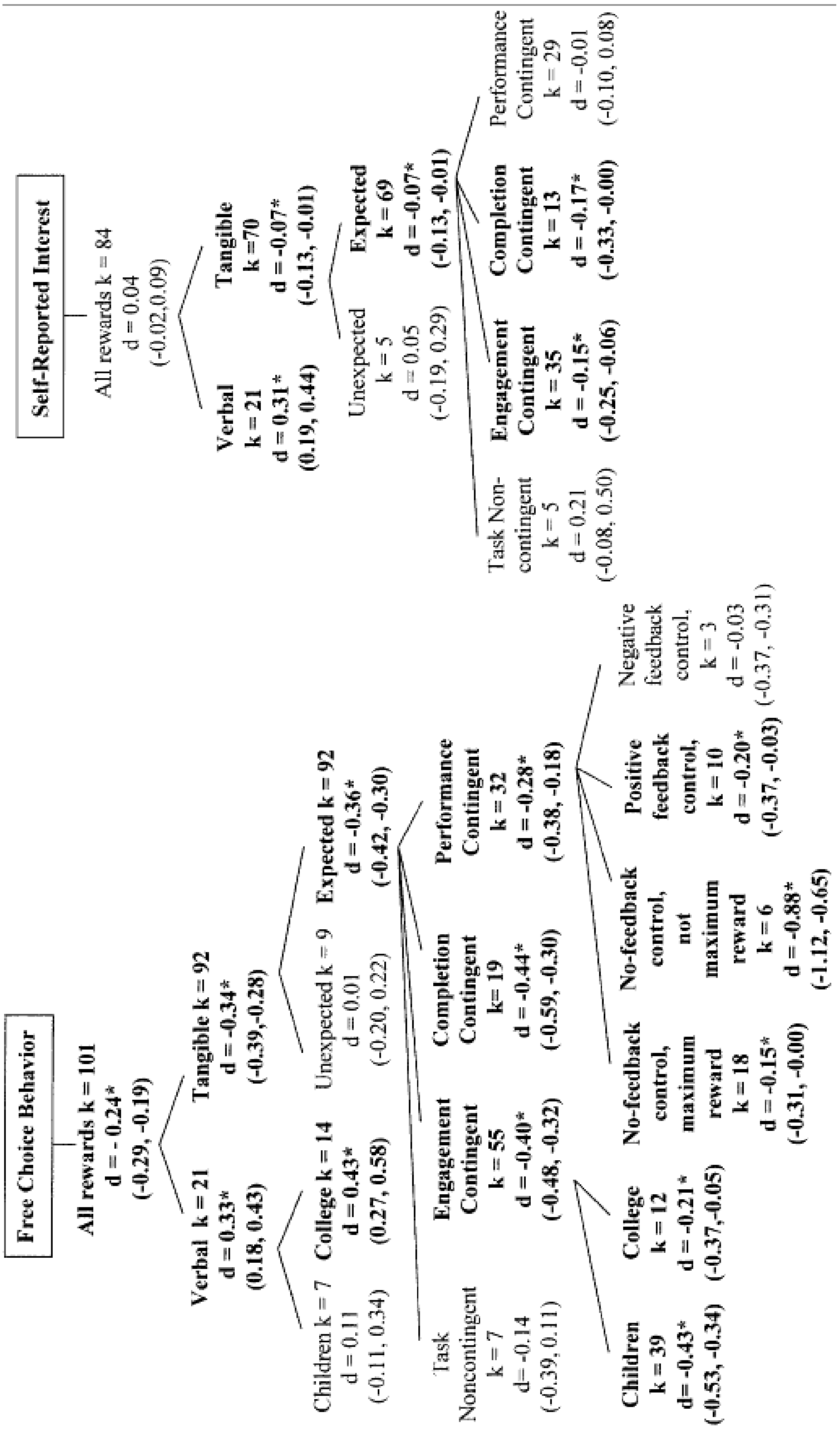
Stenzel, R., & Kübler, D. (2005). Courtersy and Idleness: Gender differences in team work and team competition. *IZA discussion paper 1768* .

Wayne, A., & Youngs, P. (2003). Teacher Characteristics and Student Achievements Gains: A Review. *Review of Educational Research , 73* (1), 89-122.

Woessmann, L. (2010). Cross-Country Evidence on Teacher Performance Pay. *IZA Discussion paper No. 5101* .

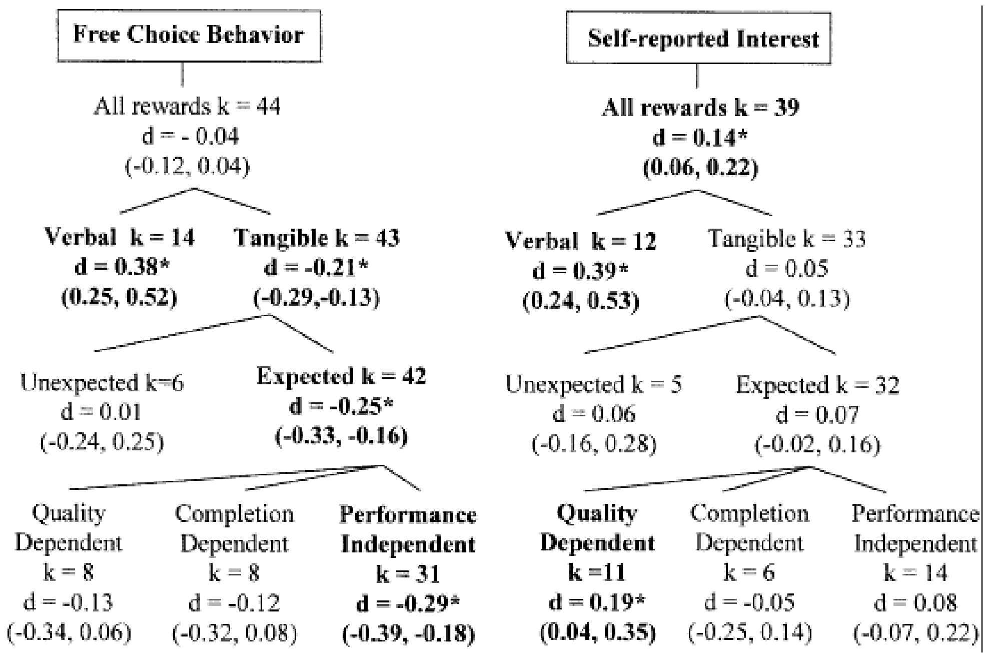
APPENDIX A1

Summary statistics Deci et al. (1999)



Appendix A2

Summary statistics Eisenberger (1996)



Appendix B: Causal interference[29]

This appendix will shortly discuss the problem of causal interference and how researchers try to deal with this problem. This will, hopefully, help the unfamiliar reader to better understand the results of the empirical studies.

B.1. Causal interference

When programs, like introducing a PRP scheme, are studied the main idea is to estimate causal effects of the program/treatment. Yet estimating these causal effects is not that easy. Let us first define the term causal effect. Suppose a certain individual is either exposed to some treatment (T) or to some control treatment (C). Let the treatment be introduced at time $t_1$and finished at time t_2. At t_2 the dependent variable Y,which the program tries to influence, is measured. The causal effect of the treatment T on variable Y can then be defined as follows:

Let $Y(T)$ be the value of Y measured at t_2 given that the individual received the treatment T.
Let $Y(C)$ be the value of Y measured at t_2 given that the individual received the control treatment C.
Then $Y(T) - Y(C)$ is the causal effect of the treatment T versus control treatment C on Y for that particular individual between time t_1, t_2.

For example, the treatment T is a PRP pay scheme for teachers and the control treatment C is a regular pay scheme. There exist only one student which at the end of the school year gets a grade, i.e. the depend variable Y. Suppose that if the student was assigned a teacher under a PRP scheme at time t_1, his grade at the end of the school year, at time t_2, is a 8; and suppose that if that *same* student instead was assigned a regular teachers at time t_1, his grade would have been a 7 at time t_2. Then the causal effect on the test score for that particular student between time t_1 and t_2 of the PRP scheme versus the regular pay scheme is $8 - 7 = 1$ point higher scored. Thus to determine the causal effect we need to answers two questions:

[29] This chapter is based on the work of Rubin (1974) and Duflo et al. (2007).

How will the student score when under the treatment $(Y(T))$? and how will the student score, that was under the treatment, if he had not received the treatment $(Y(C))$? (the counterfactual question)

As the reader might have deducted from the above questions there is a problem. As it is impossible to answer those two question at the same time. In other words it is impossible to measure both $Y(T)$ and $Y(C)$ for the *same* individual. A student was either assigned to a teacher with a PRP pay scheme or to a teacher with a regular pay scheme, but cannot be assigned to two teachers at the same time. It is therefore impossible to determine individual treatments effects. However we can estimate the average impact of the program on a group by introducing a comparison group that is not exposed to the treatment.

For simplicity assume for now there are only two individuals. One individual is exposed to the treatment and one who is exposed to the control. The causal effect than becomes the average of the two, i.e.:

$$\frac{1}{2}[Y_1^T - Y_1^C + Y_2^T - Y_2^C] \qquad [1]^{30}$$

As we have seen in reality this estimate is either

$$\frac{1}{2}[Y_1^T - Y_2^C] \qquad [2]$$

Or

$$\frac{1}{2}[Y_2^T - Y_1^C] \qquad [3]$$

As we cannot observe $Y_i^T$and Y_i^C for the same individual. Equation [2] and [3] are by no means equal to equation [1], as the comparison group might not be the same as the treatment group. If this is the case the impact that is found can be attributed both to the treatment effect as well as to pre-existing differences (i.e. 'the selection bias'). As we cannot differentiate between the two we cannot reliable estimate the treatment effects. Let us continue our numerical example were $Y_1^T = 8$, $Y_1^C = 7$, $Y_2^T = 8$ and $Y_2^C = 6$. In this case had the treatment

[30] Which might be generalized to: $\frac{1}{N}\sum_{i=1}^{N}\left[Y_i^T - Y_i^C\right]$ where $i\{1,2,\dots,N\}$

group not been treated it would have scored higher than the control group. Our causal effect estimator would become:

$$\frac{1}{2}[8-6+8-7]=1.5$$

Our results are 0.5 point higher and contain a selection bias of 0.5 point.

Luckily there are several ways to deal with this selection bias problem. Namely, randomization, controlling for observables, regression discontinuity design and difference-in-difference estimates.

B.1.1. Randomization:

One way to remove the selection bias is by randomly assign the individuals to a treatment and comparison group. As the individuals are randomly assigned they differ, in expectations, only in their exposure to the treatment. Had neither randomized groups been exposed to the treatment they would have had the same outcome. Naturally, larger samples that have smaller variance better predict the 'true' average causal effect than smaller samples. Randomization thus provides a manner to assume that $Y_1^T = Y_2^T$ and $Y_1^C = Y_2^C$.

B.1.2. Controlling for Observables:

Suppose that the experimenter, on basis of some 'special' information, knows that the only relevant differences between the treatment and control group is caused by a set of variables X. Thus when the outcome controls for this set of variables X the expected outcome between treatment and control group is the same. For this to be true, however, an important assumption has to be made. Namely, that the set of variables X indeed contain all relevant variables, i.e. there is no omitted variable. Sadly, this assumption is not testable and a skeptical observer can always make the argument that some variable that is not included influences the outcome. Therefore the experimenter can only argue on theoretical grounds that he has included all relevant variables. Controlling for a set of observables can be done via non-parametric matching, propensity scores or via a regression framework. A special case of controlling for observables is via a regression discontinuity design.

B.1.3. Difference-in-Difference estimates

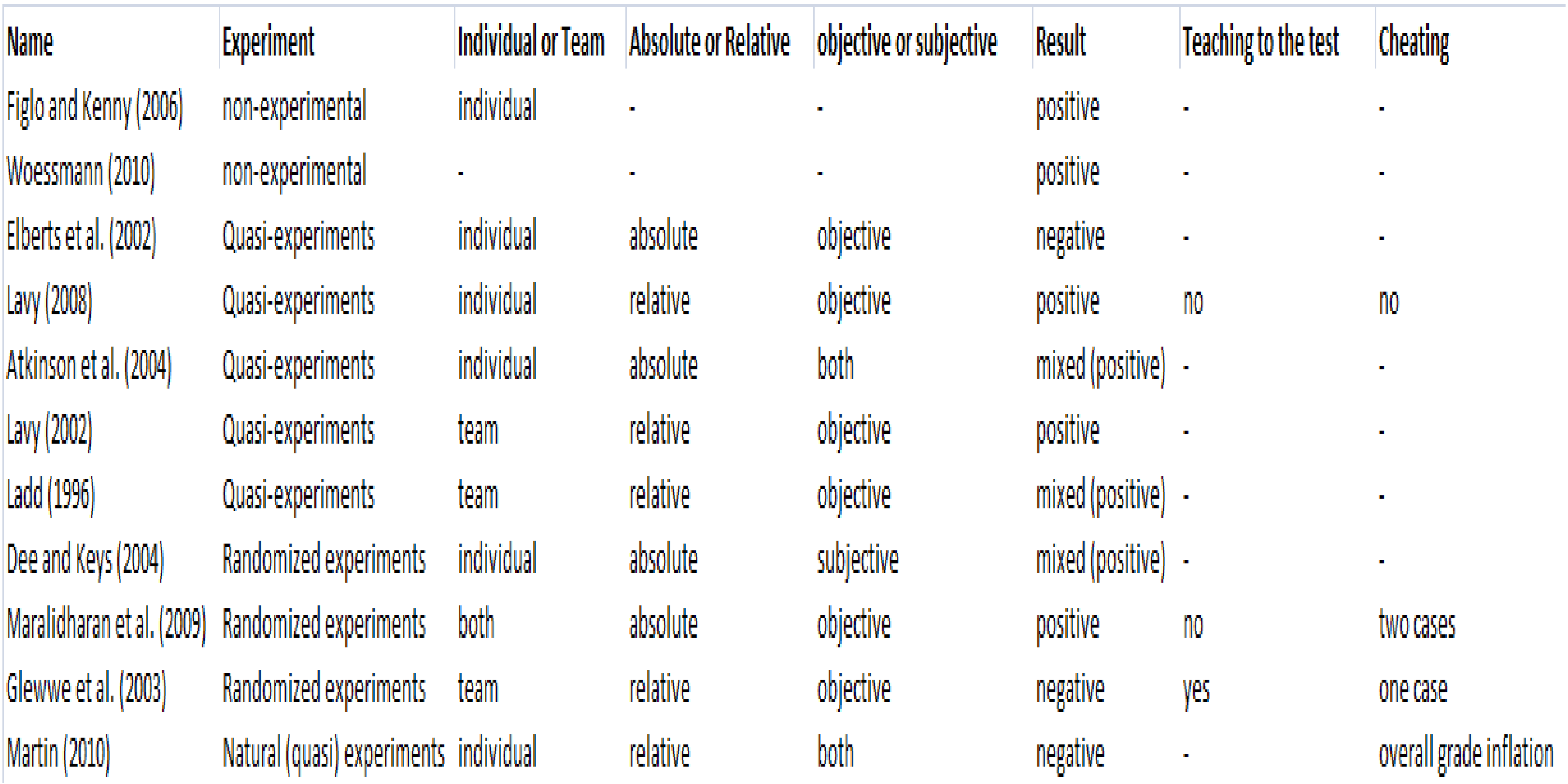
Difference-in-difference estimates make use of pre-period differences in outcomes between the treatment group and the control group. Under the assumption that these difference will remain constant if the treatment group would not have been treated, i.e. the two groups had followed parallel trends, it can successfully control for the selection bias.

Appendix C

Secundair education	Agree (%)	Neutral (%)	Disagree (%)
	20-30 vs. 50-60	20-30 vs. 50-60	20-30 vs. 50-60
42.A There are good criteria available on which teachers can be assessed if they are entitled to a higher pay scale	36.5 - 41.0	25.5 - 24.3	38.0 - 34.7
42.B Principals are capable to assess teachers transparently	16.4 - 14.0	34.3 - 20.3	**49.3 - 65.7**
42.C PRP incentivize teachers to perform better	**41.4 - 23.0**	22.9 - 23.6	35.7 - 53.4
42.D With PRP teachers get a better career perspective	**45.7 - 27.1**	23.2 - 23.2	31.2 - 49.7
42.E PRP is more fair as additional effort and/or quality is rewarded	**59.6 - 36.4**	16.3 - 19.7	24.1 - 43.9
42.F PRP can be used just as well in education as in other sectors.	**41.4 - 31.7**	19.3 - 14.7	39.3 - 53.6
42.G PRP causes teachers to be rewarded on basis of wrong criteria	**35.5-54.8**	33.3 - 22.2	31.2 - 23.0
42.H PRP causes division among teachers	67.9 - 72.6	17.5 - 18.1	14.6 - 9.3
42.I PRP brings too much power to school principals	**60.6 - 79.0**	31.1 - 13.7	8.3 - 7.3
41. Do you think that the salary scale should become depended on the individual performance of the teacher?	34.9-40.7	22.5-20.2	42.6-39.1

Primary Education	**Agree (%)**	**Neutral (%)**	**Disagree (%)**
	20-30 vs. 50-60	20-30 vs. 50-60	20-30 vs. 50-60
43.A There are good criteria available on which teachers can be assessed if they are entitled to a higher pay scale	40.7 - 38.7	17.6 - 21.6	41.7 - 39.8
43.B Principals are capable to assess teachers transparently	**31.2 - 23.7**	20.4 - 24.4	48.4 - 51.9
43.C PRP incentivize teachers to perform better	**45.9 - 23.5**	21.0 - 18.6	33. - 57.9
43.D With PRP teachers get a better career perspective	**55.8 - 40.4**	16.0 - 19.0	28.2 - 40.6
43.E PRP is more fair as additional effort and/or quality is rewarded	**57.1 - 43.3**	15.1 - 17.6	27.8 - 39.1
43.F PRP can be used just as well in education as in other sectors.	**46.0 - 37.8**	13.8 - 15.6	40.2 - 46.5
43.G PRP causes teachers to be rewarded on basis of wrong criteria	**33.7-46.6**	31.0 - 27.8	35.4 - 25.5
43.H PRP causes division among teachers	67.0 - 70.4	18.2 - 16.8	14.8 - 12.8
43.I PRP brings too much power to school principals	**53.6 - 64.1**	26.3 - 18.2	20.1 - 17.7
42. Do you think that the salary scale should become depended on the individual performance of the teacher?	**56.2-41.6**	-	43.8-58.4

Appendix D



Name	Experiment	Individual or Team	Absolute or Relative	objective or subjective	Result	Teaching to the test	Cheating
Figlo and Kenny (2006)	non-experimental	individual	-	-	positive	-	-
Woessmann (2010)	non-experimental	-	-	-	positive	-	-
Elberts et al. (2002)	Quasi-experiments	individual	absolute	objective	negative	-	-
Lavy (2008)	Quasi-experiments	individual	relative	objective	positive	no	no
Atkinson et al. (2004)	Quasi-experiments	individual	absolute	both	mixed (positive)	-	-
Lavy (2002)	Quasi-experiments	team	relative	objective	positive	-	-
Ladd (1996)	Quasi-experiments	team	relative	objective	mixed (positive)	-	-
Dee and Keys (2004)	Randomized experiments	individual	absolute	subjective	mixed (positive)	-	-
Maralidharan et al. (2009)	Randomized experiments	both	absolute	objective	positive	no	two cases
Glewwe et al. (2003)	Randomized experiments	team	relative	objective	negative	yes	one case
Martin (2010)	Natural (quasi) experiments	individual	relative	both	negative	-	overall grade inflation