Contract design and performance of railway maintenance: effects of incentive intensity and performance incentive schemes

Kristofer Odolinski

The Swedish National Road and Transport Research Institute (VTI), Department of Transport Economics, Box 55685, 102 15 Stockholm, Sweden (kristofer.odolinski@vti.se)

Abstract

In this paper we study the effect of contract design on the performance of railway maintenance in Sweden, using a panel data set over the period 2003-2013. The effect of incentive intensity is estimated, showing that the power of incentive schemes improve performance as measured by the number of infrastructure failures. In addition, we show that the structure of the performance incentive schemes has resulted in a reallocation of effort from failures not causing train delays to failures causing train delays, with a substantial increase in the former type of failures. This signals a deteriorating asset condition, which highlights the need to consider the long-term effects of this incentive structure. Overall, this work shows that the design of the incentive structures has a large impact on the performance of maintenance, and that the estimated effects are important to consider when assessing contract designs within this field.

Keywords: contract design, incentive intensity, maintenance, rail, infrastructure

1.0 Introduction

Government agencies often procure goods and services instead of producing it in-house. This procurement accounts for a significant part of national economies, with estimates at 12 per cent of the gross domestic product (GDP) in OECD countries (OECD 2017). Cutting costs and improving quality are frequently stated goals when introducing competitive tendering and contracting of services previously offered by a state-owned monopoly. However, careful contract design is required in order to achieve the goals of such reform, with appropriate specification and monitoring of quality along with incentive schemes to deal with moral hazard and adverse selection. Whether or not different contract designs have the desired effects needs to be tested empirically, both for policy reasons and to assess if theoretic arguments for certain designs are valid in the current case.

This paper contributes to this line of research by studying the incentive structures in railway maintenance contracts in Sweden, and their impact on the frequency of infrastructure failures. More specifically, the purpose with this paper is to provide empirical evidence on the effect of incentive intensity on the frequency of failures as well as the effect of tilted performance incentive schemes, where the latter differentiate between failures causing train delays and failures not causing train delays.

Sweden chose to gradually expose its maintenance of railways to competitive tendering in 2002. One objective of the transfer from in-house to tendered production of rail maintenance was to provide scope for innovation (Banverket 2000). To do so, firms (contractors) are given degrees of freedom by the contracts: most of the maintenance contracts are said to be outcome or performance based,[1] meaning that the contractor is not told exactly which (or the level of) activities that are to be carried out. A fixed payment is received by the contractor who needs to

[1] The formulation "…said to be…" is used in view of the extensive reference to regulations and provisions in the contracts.

meet a set of requirements with respect to the quality of maintenance. The purpose is to give the contractor an incentive to develop the maintenance production. We are therefore in a second-best situation where the client (the IM) can (and has in this case chosen to) observe the outcome rather than prescribing the input. This can, however, create a moral hazard situation as the contractor's actions may not be optimal for the client. In addition, the contractor can obtain a higher rent when information about its efficiency (technology) is not known to the client, which is the problem of adverse selection. This asymmetry in information means that the client has to make a trade-off between inducing effort (via fixed price contracts) and extracting rent from the contractor (via cost-plus contracts); the power of the incentive scheme is a central parameter in this trade-off (Laffont and Tirole, 1986).

The maintenance contracts in Sweden have an incentive scheme with a varying power, which comes in the form of different reimbursement rules for the contractors with respect to infrastructure failures. In short, the fixed payment for the (expected number of) activities required when an infrastructure failure occurs is not completely fixed due to the reimbursement rule. Depending on the level used for this rule, the contractors are either reimbursed for a large share of the cost of rectifying a failure, or they need to cover most of their expenses from the fixed payment (this is further described in section 2). Importantly, differences in this reimbursement rule imply varying incentives to prevent infrastructure failures from occurring. When designing the future maintenance contracts, it is useful to know whether this incentive scheme has an effect or not, and if there is an effect: what is the impact of a change in the power of the incentive scheme?

The fixed payment to the contractors is also connected to a performance incentive scheme, in which the contractor receives an award and/or penalty for its performance. Here we can note that a contractor will make a trade-off between different tasks within a project if these are rewarded differently and the tasks are substitutes; see for example the seminal paper by

Holmström and Milgrom (1991). Indeed, the performance incentive schemes in the maintenance contracts in Sweden are tilted, which can affect the attention to different tasks and consequently the outcome of the project. Considering that train delays are costly to society, this tilted structure of the performance incentive scheme is expected. However, one needs to ask whether it is has the desired effect or not, and what are the costs and benefits of the effect? Estimating the performance scheme's impact on the number of train delaying failures viz-à-viz other failures (not causing train delays) is a first important step in this analysis.

The theoretic work on contracts and information asymmetry in the principal-agent framework is extensive (for textbook treatments, see for example Laffont and Tirole, 1993, Laffont and Martimort, 2002 and Salanié, 2005). Wunsch (1994) is an early example of an empirical study on contract design within the field of procurement and regulation, where menus of linear contracts are calibrated for transit firms. Gagnepain and Ivaldi (2002) study the regulatory schemes for French urban transport and compare these to the optimal policies, while Roy and Yvrande-Billon (2007) use the same study object (in a different time period) to estimate differences in technical efficiency between regulatory schemes and fixed-price and cost-plus contracts. Other examples within the transport field are the study by Dalen and Gómez-Lobo (2003) - showing that high-powered incentive schemes reduce operating costs for bus companies in Norway - and the study by Piacenza (2006) with similar results for Italian public transport.

To the author's knowledge, an econometric test of the effect of incentive intensity has not been made in field of rail infrastructure management. Nonetheless, Vickerman (2004) provides an exploration of incentives in transport infrastructure maintenance, and a case study on incentives in rail maintenance contracts is made by Stenbeck (2008). Moreover, studies on the power of incentive schemes in procurement and regulation usually compare different types of contracts (for example fixed-price contracts compared to cost-plus contracts). We can however make use of the variation in the incentive intensity in the cost-reimbursement contracts that are used for railway maintenance services in Sweden. This enables an estimation of the effect of incentive intensity within the same contract type.

There is a wide literature on the effects of performance payments; see for example Lazear and Oyer (2013) for a review of theories and empirical findings on incentives and performance (among other topics) in personnel economics.[2] A recent study on procurement and performance incentives is made by Lewis and Bajari (2014), showing that penalties induced effort in high-way construction contracts (with welfare improvements and low contractor costs according to simulations). Our study adds to this literature by estimating the effects of performance incentives in rail infrastructure management, focusing on the reallocation of efforts from one type of failure to another.

The outline of the paper is as follows. Section 2 describes the main ingredients of the railway maintenance contracts, and the related research questions of the paper. A description of the data is provided in section 3. The models to be estimated and the estimation method are presented in section 4. The results are presented in section 5, followed by a discussion of our findings in section 6. Section 7 concludes.

2.0. Maintenance contract design and research questions

Most of the railway maintenance contracts in Sweden are performance-based contracts. These contracts are a mix between a fixed price and a cost-plus contract, i.e. a fixed payment is received for certain activities while others have variable payments.

Most contracts have a fixed payment for the (expected number of) activities required when an infrastructure failure occurs. However, the cost for each activity is capped; a clause states

[2] Other examples are Rosenthal and Frank (2006) and van Herck et al. (2010) who provide reviews of empirical evidence in the health sector, Podgursky and Springer (2007) present evidence in the education sector and Devers et al. (2007) is a review of evidence on executive pay and firm performance.

when the cost of rectifying a failure is included in the fixed payment to the contractor. It also indicates that when the cost is higher than the cap, the contractor is paid according to the variable cost for the amount above the specified cost level. For intuition, consider the following example illustrated in Figure 1: the contractor receives a fixed payment for rectifying failures during one year. A clause states that if the contractor's cost of rectifying one failure is above SEK 10 000, the contractor will be reimbursed the amount above SEK 10 000. Hence, if the total cost of rectifying one failure is SEK 15 000, the contractor will be reimbursed SEK 5 000. A higher cost-cap in Figure 1 implies that the contractor is reimbursed less. This reimbursement rule varies between contracts, creating different levels of power in the incentives. The same type of reimbursement rule is used for maintenance activities that prevent infrastructure failures, i.e. for fixing a defect before it becomes a failure.

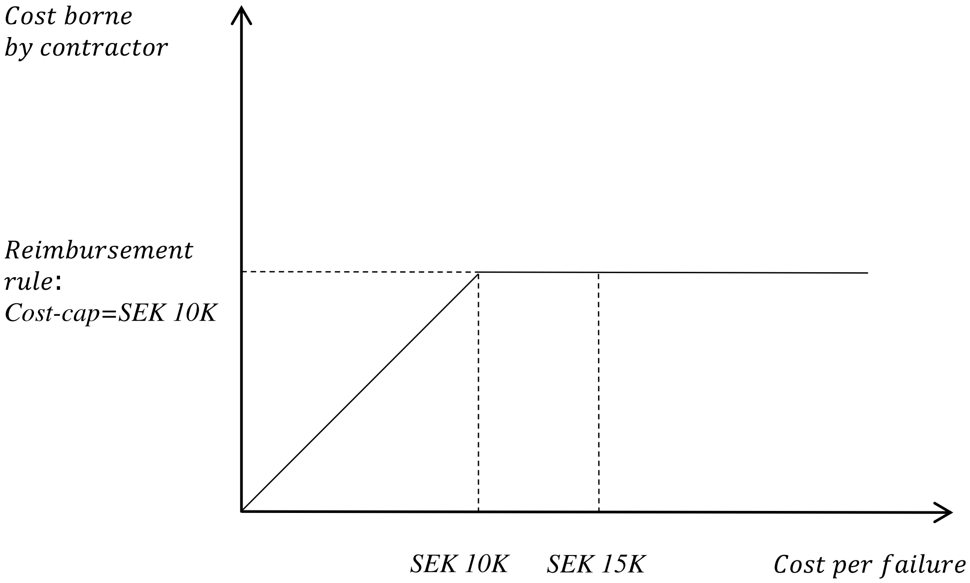

Figure 1 – Illustration of the reimbursement rule

Apart from capping the contractor's cost for some activities, the contracts also include a bonus and/or penalty linked to the number of failures in the maintenance area. These are tilted towards failures that cause train delays, which imply that an average train delay failure will have a larger impact on the bonus or penalty compared to an average failure not causing a train delay. For example, a contract using a performance index has the weight 1.8 for train delays while the weight is 1 for other failures and 0.2 for a measure of track geometry. Other contracts have bonuses and/or penalties linked to target values for train delays while failures not causing a train delay are excluded.[3] In summary, the contracts are designed so that a contractor prefers a failure that is not causing train delays instead of a train delay failure. This will tilt the contractor's maintenance strategy towards preventing failures causing train delays. For example, consider a situation where two defects are found that have the same expected corrective maintenance cost, but one defect is more likely to cause train delays than the other (which can be due to the type or the severity of the defect). The contractor will then benefit more by first fixing the defect that is more likely to cause train delays, which increases the probability of the other defect to become a failure. Still, the number of other failures must be handled in order to cap the risk of them causing train delays (for example, fixing failures require time slots on the tracks, and trains will eventually need to be rescheduled when the number of failures grows), which means that not all effort can be allocated to the prevention of train delaying failures.

In this paper, performance refers to the number of infrastructure failures that needs to be fixed immediately or within two weeks. There are other types of infrastructure quality indicators such as "minor" deviations in track geometry or other defects that require a preventive maintenance, i.e. maintenance that prevent infrastructure failures. Thus, a lack of preventive maintenance will result in a failure that requires corrective maintenance.

Given the design of the contracts, we formulate the following research questions:

[3] There are also contracts that do not have any bonus connected to train delays. However, all procured maintenance contracts have a penalty for the contractor if a time limit to rectify a train delay failure is exceeded. For example, the penalty can be SEK 10 000 if it takes more than five hours to rectify a train delay failure.

1. Do variations in the reimbursement rule affect the performance of maintenance contracts, measured as the frequency of failures?
2. Do performance incentive schemes tilted against train delays have an effect on the relationship between the number of failures causing train delays and other failures?

Below we present the dataset available for the analysis, followed by a presentation of the models to be estimated in order to answer the research questions and provide empirical evidence on the impact of the incentive structures.

3.0 Data

The Swedish railway network comprises about 16 500 track km, of which about 14 000 km is managed by the Swedish Transport Administration (which we hereafter refer to as the infrastructure manager, IM). The railway network is divided into 35 contract areas. The IM has provided copies of the signed contracts for these areas, which include information on the incentive structures with respect to infrastructure failures. The available information covers the period 2003-2013, during which maintenance has been gradually exposed to competitive tendering (starting in 2002 and as of 2013, 95 per cent of the railway network had been tendered in competition).

Information on the number of infrastructure failures has also been collected from the IM. The data constitute all failures reported to the IM that needed to be fixed immediately or within two weeks, which is the Swedish IM's definition of a failure. It is indicated in the data whether a failure has caused a train delay or not. Reports of failures come from the train management system and may emanate from operators, train drivers, maintenance personnel as well as the public. There are many different causes of failures. Some are strictly exogenous with respect to maintenance such as animals or humans hit by a train, sabotage etc. These failures are out of the contractors' control and are not included in the analysis, meaning that only failures occurring because of deterioration and/or poor maintenance of the infrastructure are analysed.

Infrastructure characteristics such as track length, rail weight and quality class are part of our dataset, as well as traffic volume. These are important factors to control for when analysing the impact contract designs have on the frequency of failures. For example, rail weight is a quality indicator, where heavier rail is more resilient towards the damage mechanisms causing track deterioration (failures), while the quality class is linked to differences in linespeed and requirements on track geometry standards. The traffic information collected is gross ton density (ton-km/route-km), which from an engineering perspective is a key driver of wear and tear, i.e. infrastructure failures.

Table 1 shows the descriptive statistics for failures and the explanatory variables included in the estimations. The level of detail varies between our variables. The failures are reported to the IM at a very disaggregate level, with information on which station (or between which two stations) the failure was located. These parts of the tracks are called segments. Some of the segments have a short track length (for example 10 metres) as they only constitute a switch or a bridge, while some segments comprise several km of track. Information on rail weight and quality class is also reported for each segment. However, traffic volume is available at a more aggregate level, defined as track sections that on average include about 11 segments, while each contract area on average comprise about 6 track sections. The length relationship between a segment, track section and contract area is thus $segment\ km < section\ km < contract\ area\ km$.

Table 1 - Data 2003-2013

Variable (24 940 obs.)	Mean	Std. Dev.	Min	Max
SEG.: Failures, total	6.32	17.75	0	482
SEG.: Failures, train delays	1.11	2.85	0	101
SEG.: Failures, not train delays	5.22	15.47	0	413
SEG.: Track length, metres	5 871	5 695	10	43 870
SEG.: Rail weight, kg	51.58	5.68	27	63
SEG.: Quality class, 0-5*	2.06	1.26	0	5
SEC.: Traffic density, million gross tonnes**	8.29	8.62	0.00	49.79
ARE.: Year tendered in competition, dummy	0.58	0.49	0	1
ARE.: Mix between tend. and not tend. in comp., dummy	0.08	0.27	0	1
Subset of data used for research question 1 (8 528 obs.)				
ARE.: Cost-cap, thousand SEK	7.66	4.04	5	20
SEG.: Failures, total	5.95	15.11	0	482
SEG.: Failures, train delays	1.12	2.85	0	101
SEG.: Failures, not train delays	4.83	12.91	0	381
SEG.: Track length, metres	6 052	6 262	10	43 077
SEG.: Rail weight, kg	51.73	5.82	32	60
SEG.: Quality class, 0-5*	2.01	1.32	0	5
SEC.: Traffic density, million gross tonnes**	7.89	9.32	0.00	49.79

SEG = Information available for segments, SEC = Information available for sections, ARE = Information available for contract areas, *A high value implies a low speed line with less strict requirement on track geometry standards compared to a high-speed line (Banverket 1997), **Traffic density = (Million gross ton-km/Route km)

In total, we have 24 940 observations on the segments administered by the IM over the period 2003-2013. However, only tendered contracts have a reimbursement rule (described in section 2). This information is available for a third of the observations, with descriptive statistics in Table 1, indicating that the cost-cap (reimbursement rule) varies between SEK 5 000 and SEK 20 000 per failure.[4] Out of the 32 contract areas that are part of the dataset, five have been subject to a change in the cost-cap. The changes comprise increases in the cost-cap from SEK

[4] As an indication of what is required to reach the cost-caps, we consider a labour cost at around SEK 1000 per hour (and exclude costs for material). The lowest cost cap (SEK 5000) is then reached after 5 person hours, while the highest cost-cap requires 20 person hours. Indeed, minor failures can be solved without reaching the cost-cap, while more severe failures, requiring more person hours and material, will quickly reach the cost-cap.

5 000 to SEK 7 000, from SEK 5 000 to SEK 10 000, from SEK 8 000 to SEK 10 000, and from SEK 15 000 to SEK 18 000.

For the tendered contracts, we note that the average number of failures per segment and year is 5.95, while the average for a contract is 354.84 failures per year. Specifically, we use this subset of the data to answer research question 1. The entire dataset can be used for research question 2. The reason is that the number of failures prior to competitive tendering are used to evaluate the effect of the performance incentive schemes, which are tilted towards failures causing train delays (see description of models and estimation method below). Here we can note that there are on average 2505 failures per year causing train delays, while the average per segment and year is 1.11 (as indicated in Table 1). The corresponding figures for failures not causing train delays is 14337 per year and 5.22 per segment and year.

During 2003-2009 a train had to be delayed more than 5 minutes between two stations for a failure causing the delay to be reported as such. This definition was changed to 3 minutes in 2010. To consistently analyse the number of train delay failures during 2003-2013, we only include failures causing more than a 5-minute delay. Failures causing less than 5 minutes of delay are therefore defined as a regular failure in this study. Furthermore, we were not able to get consistent information about the knock-on effects of a first train being delayed, meaning that it is impossible to report the total number of delay minutes per failure from the available data.

In the analysis of the reimbursement rule we exclude the contracts that are not performance based, which are primarily used for the newly built railway lines. Moreover, we exclude yards when analysing train delay failures because these are exempted from the bonus/penalty system with respect to train delays in the maintenance contracts.

4.0 Models and estimation method

We use three different models to answer our research questions. The first model considers the reimbursement rule, and the second and third models address the impact of tilted performance schemes. The total number of failures (f_{it}) is the dependent variable in the first model, and comprise failures that caused train delays (f_{it}^{T}) and other failures that did not cause train delays (f_{it}^{O}), that is $f_{it} = f_{it}^{T} + f_{it}^{O}$. The number of failures consists of non-negative discrete values, i.e. it is a count variable. We therefore use count data models, where we consider the first model to have the following (exponential) conditional mean:

$$E\left[f_{it} \,\middle|\, \sum_{k=1}^{K} x_{kit}\,, c_{it}, \alpha_i\right] = \alpha_i exp(\sum_{k=1}^{K} x_{kit}\,\beta_k + \mu c_{it}), \qquad (1)$$

where $t = 1,..,T(i)$ years, and i indicates the individual segments. α_i is segment-specific effects, and β_k is a vector of parameters for the effect of the explanatory variables $\sum_{k=1}^{K} x_{kit}$, which include variables for the infrastructure characteristics, traffic volume and year dummy variables. c_{it} is the cost-cap linked to the reimbursement rule and μ its parameter. Recall that a higher cost-cap implies that the contractor is reimbursed less for a failure that is costly to rectify. Our hypothesis is that a higher cost-cap will reduce the number of failures:

$$Hypothesis\ 1 \colon \mu < 0$$

considering that the contractor then has a stronger incentive to prevent failures from occurring, i.e. it induces effort. However, the effect of the reimbursement rule can also be due to a selection effect, with more efficient contractors being awarded contracts with a high power in the incentive scheme. From a policy perspective, varying cost caps that generate a selection effect can be beneficial for the client, as it creates active inefficient types which lowers the efficient types' ability to extract rent.[5] If it is not considered in the estimation, a selection effect would, however, result in a biased estimate on the effect the cost cap has on inducing effort. A selection effect does not seem to be present in our data, where for example two out of three contractors with the highest cost-caps (SEK 15 000 to SEK 20 0000) also have contracts with the lowest caps (SEK 5 000 to SEK 8 000), but we still consider a possible effect on our cost-cap coefficient in the estimation of our model (see section 4.1).

The number of failures causing train delays (f_{it}^{T}) is the dependent variable in the second model, with a conditional mean expressed as

$$E\left[f_{it}^{T} \mid \sum_{k=1}^{K} x_{kit}\, , D_{it}, \alpha_i\right] = \alpha_i exp(\sum_{k=1}^{K} x_{kit}\, \beta_k + \vartheta_T D_{it}) \qquad (2)$$

while the number of other failures (not causing train delays; f_{it}^{O}) is the dependent variable in the third model

$$E\left[f_{it}^{O} \mid \sum_{k=1}^{K} x_{kit}\, , D_{it}, \alpha_i\right] = \alpha_i exp(\sum_{k=1}^{K} x_{kit}\, \beta_k + \vartheta_O D_{it}) \qquad (3)$$

To evaluate the effect of tilted performance schemes, we use a dummy variable (D_{it}) in the second and third model, indicating when a maintenance area is tendered in competition – that is, it is used as a proxy for a change in the effect of tilted performance incentive schemes. The reason for this estimation approach is that there is no point in time when performance incentive schemes were introduced. For example, there are examples of performance clauses in contracts

[5] As Laffont and Tirole (1993, p. 71) writes: "one could conclude that high-powered incentive schemes (…) are better because they induce better performance. While the second statement is correct, the first ignores the desirability of rent extraction. Optimal screening of the firm's technology yields over the sample good performances and high-powered schemes together with poor performances and low-powered schemes."

awarded to the in-house production unit prior to the introduction of competitive tendering. It is reasonable to assume that in-house production in general had some sort of incentive structure to reduce train delays. We can however use the *sampling benefit* from competitive tendering, which imply that it is more likely that the chosen contractor is efficient (see for example Armstrong and Sappington 2007, chapter 4). This is also suggested by the results in Odolinski and Smith (2016), showing that competitive tendering reduced maintenance cost in Sweden with about 11 per cent (which of course also can be explained by other factors than just the sampling benefit). Put differently, the tilted performance schemes induce effort towards failures causing train delays, and a more efficient contractor will thus use relatively more effort than the less efficient contractor. Our hypothesis is therefore that the use of competitive tendering will increase the effect of the tilted performance incentive schemes. The parameters ϑ_T and ϑ_O in equations 2 and 3, respectively, should differ. Hence, we state the following hypothesis:

$Hypothesis\ 2\colon \vartheta_T < \vartheta_O$

4.1 Selection bias

In *Model 1*, we include dummy variables for contractors to test if there is a selection effect (efficient contractors choosing contracts with high cost-caps) that has an impact on our coefficient for incentive intensity (μ). Moreover, there might be a selection bias in Models *2* and *3*. Specifically, the maintenance of the Swedish railway network was gradually put out to tender, with the first contract tendered in 2002 and the last part of the network tendered in competition in 2014. The estimates from the tendering dummy variables in *Model 2* and *3* will be biased if there are systematic differences between areas tendered first and tendered later that are not controlled by the independent variables; omitted variable bias will be present. A selection bias can also be present if we have reverse causality; if areas tendered first were tendered because they had high (low) probability of certain failures to occur. This issue is also addressed in Odolinski and Smith (2016), Domberger et al. (1987) and Smith and Wheat (2012). We use the same approach and include a vector of dummy variables in the estimations:

$$z_{kit} = [D_{ki}, D_{kit}, D_{Mi}, D_Y; \boldsymbol{\vartheta}] \tag{4}$$

where $k = C$ indicate when a segment is tendered in competition, $k = F$ when tendered during 2002-2004 for the first time and $k = L$ when tendered during 2005-2013 for the first time. The time period before tendered in competition is indicated by $t = B$ and $t = O$ when tendered in competition and onwards. The dummy variable D_{Mi} is used for the year when the transition from not tendered to tendered takes place, i.e. the first year an area is tendered. z_{kit} also includes year dummies ($k = Y = 2004, \dots, 2013$). $\boldsymbol{\vartheta}$ are parameters to be estimated.

As a robustness test of *Model 2* and *Model 3*, we estimate ϑ_{FBi}, ϑ_{FOi} and ϑ_{LOi} and test if $\vartheta_{FBi} = 0$, which would imply that, before the areas were tendered, we have no systematic difference between areas tendered first (in 2002-2004) compared to areas tendered later (in 2005-2013) and areas not tendered during 2003-2013.[6]

4.2 Regression model

We use count data regression models with the conditional means expressed in eq. (1)-(3).[7] First, we note that we have overdispersion in our data (variance greater than the mean), which can be explained by a large fraction of the observations having a zero value, as indicated by Figure 2

[6] The definition of areas tendered first is arbitrary because the exposure to competition was gradual, and we therefore perform sensitivity tests with respect to this definition.

[7] See for example Hausman et al. (1984) or Hilbe (2011) for specifications of the log-likelihood functions for the Poisson and the negative binomial models.

below.[8] In this case, the negative binomial model is a useful regression model, in which the conditional mean is not equal to the conditional variance. We therefore estimate a negative binomial regression model on a panel data set stretching from 2003-2013

$$\Pr\left(f_{it} \,\middle|\, \sum_{k=1}^{K} x_{kit}, \alpha_i, \delta_i\right) \qquad (5)$$

where f_{it} is the count of failures, i = track segment $1,2,\ldots,N$ and t = year $1,2\ldots,T(i)$. α_i is the individual effect as specified in (6) and δ_i is the dispersion parameter in the model, where it is assumed that $1/(1+\delta_i) \sim Beta(r,s)$. $\sum_{k=1}^{K} x_{kit}$ is a vector of k explanatory variables, including the cost cap variable in *Model 1*, and the policy dummy variables z_{kit} in *Models 2* and *3*. Track length is an important exposure variable in the models. We expect the coefficient for track length to not be significantly different from 1, meaning that a segment with track length 2 km is twice as likely to have a failure as a segment with track length 1 km, *ceteris paribus*.

We use the negative binomial random effects model, considering that Allison and Waterman (2002) found that the negative binomial model with conditional fixed effects, proposed by Hausman et al. (1984), is not a true fixed effects estimator. The Poisson conditional fixed effects estimator (that is, without the dispersion parameter in (5)) is also considered as it relies on weaker distributional assumptions compared to the negative binomial model (Cameron and Trivedi (2005)). The Poisson model can therefore be preferred when modelling the mean, yet the negative binomial model can be preferred in predicting certain probabilities.

[8] The overdispersion in Figure 2 is estimated from the pooled negative binomial model and is significantly different from zero according to a likelihood ratio test ($chi2(1) = 3.8e + 05$).

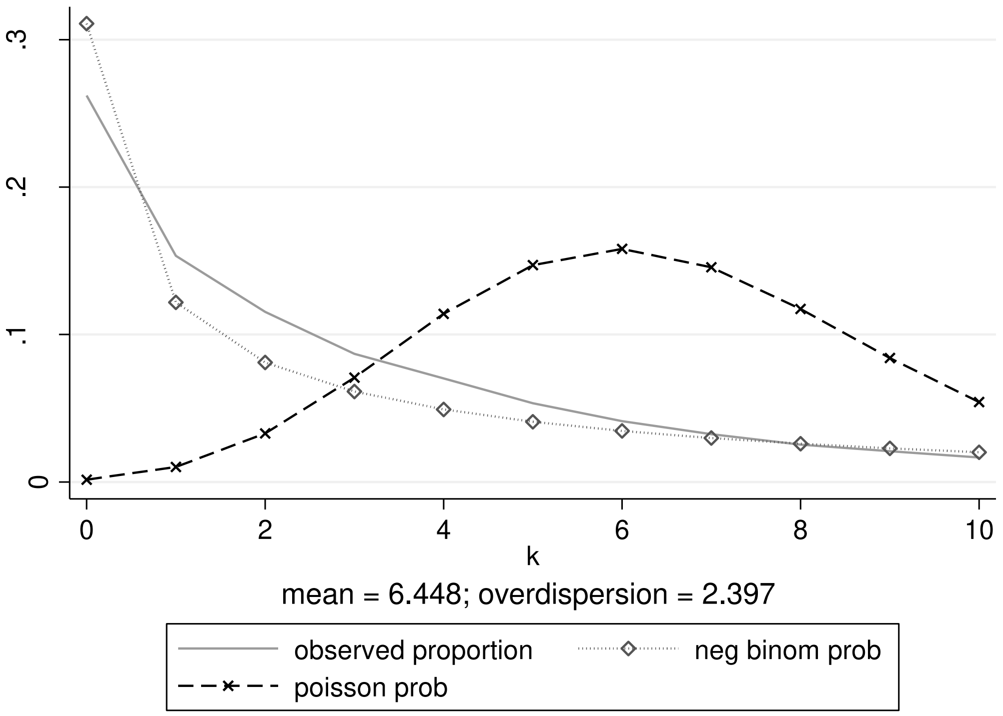

Figure 2 – Proportions of observations: observed, Poisson- and negative binomial probability

To deal with the problem of inconsistent estimates if the regressors are not independent of the individual effects α_i in the negative binomial random effect model, we use a solution first proposed by Mundlak (1978), and specify the individual specific effect as

$$\alpha_i = exp\ (\textstyle\sum_{k=1}^{K} \bar{x}_{ki}\, \gamma_k + \varepsilon_i) \qquad (6)$$

where $\bar{x}_{ki} = T^{-1} \sum_{t=1}^{T} x_{kit}$ for each $k = 1 \ldots K$ (all time-varying explanatory variables in our estimations). ε_i is unobserved heterogeneity not correlated with our regressors. Using (6) we express (1)-(3) (including the cost-cap and dummy variables in $\sum_{k=1}^{K} x_{kit}$) as

$$exp[(\textstyle\sum_{k=1}^{K} x_{kit} - \sum_{k=1}^{K} \bar{x}_{ki})\beta_k + \sum_{k=1}^{K} \bar{x}_{ki}\, \gamma_k + \varepsilon_i] \qquad (7)$$

where we control for the correlation between α_i and our regressors $\sum_{k=1}^{K} x_{kit}$ via $\sum_{k=1}^{K} \bar{x}_{ki}$. Moreover, we avoid collinearity between $\sum_{k=1}^{K} x_{kit}$ and $\sum_{k=1}^{K} \bar{x}_{ki}$ by using deviations from the mean.

A variable for different cost levels for compensation when rectifying a failure, i.e. the cost-cap linked to the reimbursement rule, is included in *Model 1.* The specification in equation (7) implies that we estimate 'within-effects' of changes in the cost-cap. Hence, we control for unobserved (time-invariant) heterogeneity between the contract areas that might explain the use of different cost-caps.

Dummy variables (z_{kit}) for competitive tendering are included in *Models 2* and *3*. Specifically, we include year dummies and a dummy variable for when a track segment belongs to a contract area tendered in competition, as well as a dummy variable indicating when there is a transition from not tendered to tendered in competition (which in most cases does not happen in the beginning of a calendar year). As we do not have a general post-tendering period (exposure to competition was gradual), we use the year dummy variables to control for general effects that occur over time, which leaves the time-specific tendering variable to pick up the impact of tendering. In line with the difference-in-differences approach, we also include a dummy variable indicating all areas tendered in competition sometime during 2003-2013 along with the time-specific tendering variable. As described in section 4.1, we also test the presence of selection bias.

5.0 Results

The model results are $\hat{\lambda}(x)_{it}$ and $\hat{\lambda}(x+1)_{it}$, which means that we estimate the expected value of failures when the explanatory variable x_{it} increases with one unit. The estimated coefficient is then $\hat{\beta} = \ln\left[\frac{\hat{\lambda}(x+1)_{it}}{\hat{\lambda}(x)_{it}}\right]$, and $e^{\hat{\beta}}$ is an incidence ratio (IRR), expressed as $\frac{\hat{\lambda}(x+1)_{it}}{\hat{\lambda}(x)_{it}}$. Hence, an

IRR<1 indicates a negative effect. The incidence ratios are reported in Tables 2 and 3 together with standard errors for the estimated coefficients $\hat{\beta}$. All estimations are carried out with Stata 12 (StataCorp.2011).

5.1 Econometric results: Model 1

Table 2 shows the results from the estimations of the first model, which include results from both the random effects model and the preferred correlated random effects model (with terms as specified in equation (7)). In the latter model, the coefficients for variables averaged over time are denoted 'between estimates' while the other coefficients are denoted 'within estimates' (referring to effects between and within segments, respectively).

Track length, which is the exposure variable, has the expected IRR of 1. The estimations also include a squared term for million gross tonne density, and the estimates reflect a non-linear relationship with the number of failures, which is shown by both the within and between estimate in the correlated random effects model. Note that only the period 2004-2013 is included in this estimation due to missing data with respect to the reimbursement rule, which means that we include year dummy variables for 2005-2013. We tested the average values over time for the year dummy variables in the estimation because we have an unbalanced panel (Wooldridge, 2013), but these were not jointly significant and dropped from the estimation.

Turning to the 'within estimate' for the cost-cap, we see that it has a negative effect on the number of failures (IRR=0.9614, p-value=0.000); we cannot reject $Hypothesis$ 1 which is related to research question 1. The incidence rate ratio at 0.9614 indicates that an increase in the cost-cap with one unit (in our case with SEK 1000) will reduce the number of failures with (100*(1-0.961)=) 3.9 percent. The average number of failures per contract and year is 355 in the sample (there are on average 59.6 segments per contract area and about 5.95 failures per segment). Hence, the estimated effect of a marginal increase in the cost-cap implies around 13.8

fewer failures per year for the average contract. Increasing the level of the cost-cap with one standard deviation (SEK 4 000), would imply about 55 fewer failures per year, which can be compared to the average of 355 failures per contract and year.

Table 2 - Results Model 1

	Random effects		*Correlated Random effects*	
	IRR	Std. Err.	IRR	Std. Err.
Constant	0.0000***	0.0000	0.0000***	0.0000
Cost-cap	0.9972	0.0050	0.9614***	0.0095
Rail weight	1.8074***	0.1278	1.5163***	0.2059
(Rail weight)^2	0.9883***	0.0013	0.9920***	0.0026
Quality class	1.0264	0.0225	1.0014	0.0431
Track length	1.0000***	0.0000	1.0000	0.0000
Million gross ton density	1.0917***	0.0076	1.0400***	0.0123
(Million gross ton density)^2	0.9971***	0.0004	0.9985***	0.0006
D.year2005	1.5806***	0.1920	1.4878***	0.1750
D.year2006	1.4409***	0.1720	1.3851***	0.1601
D.year2007	1.7515***	0.2072	1.7426***	0.1999
D.year2008	1.6040***	0.1895	1.5949***	0.1828
D.year2009	1.4920***	0.1761	1.5073***	0.1729
D.year2010	1.4200***	0.1675	1.4492***	0.1662
D.year2011	1.4428***	0.1702	1.4444***	0.1656
D.year2012	1.4443***	0.1707	1.4544***	0.1675
D.year2013	1.5331***	0.1810	1.5475***	0.1787
$T^{-1}\sum_{t=1}^{T}$ Cost- cap	-	-	1.0077	0.0057
$T^{-1}\sum_{t=1}^{T}$ Rail weight	-	-	1.8926***	0.1561
$T^{-1}\sum_{t=1}^{T}$(Rail weight)^2	-	-	0.9872	0.0016
$T^{-1}\sum_{t=1}^{T}$ Quality class	-	-	1.0344	0.0280
$T^{-1}\sum_{t=1}^{T}$ Track length	-	-	1.0001***	0.0000
$T^{-1}\sum_{t=1}^{T}$ Million gross ton density	-	-	1.1141***	0.0094
$T^{-1}\sum_{t=1}^{T}$(Million gross ton density)^2	-	-	0.9965***	0.0004

***, **, *: Significance at 1%, 5%, 10% level, respectively

Log likelihood: Random effects model= -18 844.355, Correlated Random effects model= -18 810.507

Number of observations = 8 528

We included dummy variables for contractors to test if there is a selection effect generating a biased estimate of the effect the cost-cap has on inducing effort. This did not have a significant impact on the cost-cap estimate (the IRR is then 0.9515 with p-value 0.000). Moreover, we also estimate the model using Poisson conditional fixed effects, considering that it relies on weaker distributional assumptions than the negative binomial model. The results are presented in Table 4 in the appendix, showing a similar effect of the cost-cap: The IRR is 0.9727 and statistically significant (p-value 0.033).

5.2 Econometric results: Model 2 and Model 3

The estimation results from *Model 2* and *Model 3* using correlated random effects are presented in Table 3. In *Model 2*, the number of failures causing train delays is used as the dependent variable. The dependent variable in *Model 3* is failures not causing a train delay.

The effects of rail weight, track length and traffic are similar to the results in *Model 1*. However, the IRR for quality class (a high number indicates low line speeds) is positive and statistically significant in the model for failures causing train delays. That is, these failures are more frequent on tracks with low line speeds (tracks with poor track geometry standards).

Importantly, there are differences in the effect of the competitive tendering between *Model 2* and *Model 3* according to the estimation results. In *Model 2*, the IRR for competitive tendering ("D.Year tendered in competition"), is 0.9593 (p-value=0.191) with a 95 per cent confidence interval at [0.9014, 1.0209], which indicates a negative effect on the number of failures causing train delays, yet not statistically significant. The IRR for competitive tendering in *Model 3* is 1.0696 (p-value=0.000) with a 95 per cent confidence interval at [1.0321, 1.1086], which implies that the number of failures that has not caused a train delay is increasing when tendered in competition. The lower parameter estimate in *Model 2* compared to *3,* and more importantly the non-overlapping 95 per cent confidence intervals, implies that we cannot reject $Hypothesis$ 2, which is linked to research question 2. Considering that there were on average 2505 failures causing train delays per year (1.11 per segment), the impact of the performance scheme imply (2505*(1-0.9593=) 102 fewer failures causing train delays per year (keeping in mind that the estimate was not statistically significant). This can be compared to the impact on the other failures, which corresponds to (14337*(1.0729-1)=) 998 more failures per year (based on 14337 failures per year).

Note that we include a dummy variable indicating all segments tendered sometime during 2003-2013 to control for any general feature among these segments that also apply before tendering. The corresponding IRR is above one in both models, but not statistically significant (p-values at 0.787 and 0.824 in the respective models). Moreover, no selection bias was found using the dummy variable approach described in section 4.1. That is, the parameter estimate of interest ϑ_{FBi} is not significantly different from zero (p-value=0.872 and 0.204 in *Model 2* and *3* respectively).

Results from the Poisson conditional fixed effects models are presented in the appendix, showing that the IRR for competitive tendering ("D.Year tendered in competition"), is 0.9983 and not statistically significant for failures causing train delays (p-value =0.973), while it is 1.0871 and statistically significant for other failures, which is in line with the negative binomial model results.

Table 3 - Results Model 2 and 3: Correlated Random Effects

	Model 2 - Train delay failures		*Model 3 - Other failures*	
	IRR	Std. Err.	IRR	Std. Err.
Constant	0.0000***	0.0000	0.0000***	0.0000
D.Mix tendered and not tendered in comp.	1.0132	0.0336	1.0460**	0.0196
D.Year tendered in competition	0.9593	0.0305	1.0696***	0.0195
D.If tendered in comp. in 2003-2013	1.1147	0.4478	1.0525	0.2425
Rail weight	1.4952***	0.1778	0.9775	0.0531
(Rail weight)^2	0.9919***	0.0022	1.0000	0.0010
Quality class	1.0793**	0.0401	1.0077	0.0204
Track length	1.0000	0.0000	1.0000	0.0000
Million gross ton density	1.0513***	0.0118	1.0329***	0.0070
(Million gross ton density)^2	0.9988***	0.0005	0.9985***	0.0003
D.year2004	0.8994***	0.0353	0.9148***	0.0194
D.year2005	0.9352*	0.0366	0.8503***	0.0185
D.year2006	0.9411	0.0374	0.8301***	0.0185
D.year2007	1.1399***	0.0446	1.0471**	0.0227
D.year2008	1.0517	0.0441	0.9802	0.0227
D.year2009	0.9906	0.0435	0.9802	0.0235
D.year2010	1.2507***	0.0553	0.9242***	0.0232
D.year2011	1.3211***	0.0602	0.8749***	0.0230
D.year2012	1.2209***	0.0573	0.7621***	0.0208
D.year2013	1.3843***	0.0647	0.7974***	0.0219
$T^{-1}\sum_{t=1}^{T}\text{D. Mix tend. and not tend. in comp.}$	2.6137*	1.3885	2.1869**	0.8187
$T^{-1}\sum_{t=1}^{T}\text{D. Year tendered in competition}$	0.9615	0.0967	0.9849	0.0793
$T^{-1}\sum_{t=1}^{T}\text{D. If tend. in comp. in } 2003-2013$	0.9921	0.1199	1.0354	0.1007
$T^{-1}\sum_{t=1}^{T}\text{Rail weight}$	2.9059***	0.2435	1.9981***	0.1207
$(T^{-1}\sum_{t=1}^{T}\text{Rail weight})\text{^}2$	0.9793***	0.0016	0.9863***	0.0011
$T^{-1}\sum_{t=1}^{T}\text{Quality class}$	0.9669	0.0220	1.0453**	0.0197
$T^{-1}\sum_{t=1}^{T}\text{Track length}$	1.0000***	0.0000	1.0000***	0.0000
$T^{-1}\sum_{t=1}^{T}\text{Million gross ton density}$	1.1775***	0.0092	1.1004***	0.0072
$T^{-1}\sum_{t=1}^{T}(\text{Million gross ton density})\text{^}2$	0.9939***	0.0004	0.9963***	0.0004

***, **, *: Significance at 1%, 5%, 10% level, respectively

Note: year dummy variables $T^{-1}\sum_{t=1}^{T}\text{D. year2004}$ to $T^{-1}\sum_{t=1}^{T}\text{D. year2013}$ are jointly significant and included in the estimations, but dropped from Table 3 for expositional convenience.

Log likelihood: *Model 2* = -27 527.569, *Model 3* = -51 306.639

Number of observations in *Model 2* and *Model 3*: 24 940

6.0 Discussion

The design of contracts is vital for the outcome of the maintenance projects, which places high demands on the IM as a client. The presence of hidden information and hidden action can result in inefficient outcomes if not judiciously handled. Incentive contracts linear in costs can be used to alleviate the problems incurred by these information asymmetries. Indeed, this type of contract is used by the IM in the tendering of maintenance contracts in Sweden, where different reimbursement rules (cost-caps) have been used over the years of competitive tendering creating different incentive intensities. The estimation results show that an increase in the cost-cap (higher incentive intensity as the contractor is reimbursed less) reduces the number of infrastructure failures. The estimated impact is substantial, considering that a one standard deviation increase (SEK 4000) in the cost-cap imply 55 fewer failures per year and contract area, which is about 15 per cent of the total number of failures per year and contract area. Here, we can note that all the changes made with respect to the reimbursement rule have been increases in the cost-cap. That is, the IM have chosen to increase the incentive intensity in the contracts.

Does this result imply that we should have a high cost-cap in all maintenance contracts? Not necessarily. A high cost-cap indicates that we move closer to a fixed price contract which induces effort, but this will make it easier for the efficient contractor to extract rent. Moreover, we will have a low level of competition if inefficient types do not take part in the bidding when the cost-caps are too high, with the efficient type(s) being able to extract higher rents. Still, the estimated impact of increased incentive intensity can be used in the pursuit of a desired balance between the expected number of infrastructure failures and the expected cost of a maintenance contract. The next step is thus to evaluate how different levels of the cost-caps affects maintenance costs.

There might be possibilities for the contractors to misguide the IM on the number of failures and/or the cost of solving a failure. If we assume that it is *difficult to misreport the costs* of solving a failure, a contractor might then report several failures as one failure in order to reach the cost-cap. This could imply that our cost-cap coefficient is overestimated, considering that a higher cap would require more failures to be collected in the report to the IM. On the other hand, if it is *easy to misguide the IM on the cost* of solving a failure, then there is no incentive to understate the number of failures (assuming the contractors' success in misguiding the IM is not dependent on the stated cost level per failure). The contractor could in this case choose to exaggerate the number of failures and/or overstate the cost of each failure, where the mix depends on its effectiveness. This would reduce the effect of the cost-cap, and our coefficient is then underestimated compared to a case with no misreports. Moreover, the benefit of preventing a failure is reduced if it is easy to misguide the IM on costs, which would lead to more (actual) failures. In fact, a contractor could in theory cover its cost of solving a failure (the part below the cost-cap) by stating a higher cost for higher caps, making the level of the cap irrelevant for the incentive to prevent failures. That is, the number of actual failures would not vary with the cost-cap for contractors that misguide the IM on the true costs. Again, the estimated effect of the cost-cap coefficient is then underestimated; it would be higher (result in fewer failures) if the IM implemented (more) cost monitoring measures.

An important task of the contractors is to prevent infrastructure failures that are causing train delays. A robust and reliable railway infrastructure is an objective often stated by the IM. Undeniably, this objective is reflected in the design of the maintenance contracts, with performance incentive schemes tilted towards failures causing train delays. This makes it beneficial for the contractors to focus on this group of infrastructure failures. The estimation results confirm our hypothesis, suggesting that effort is tilted towards preventing failures causing train delays at the expense of preventing other failures.

Are the performance incentive schemes beneficial with respect to the performance of the railway infrastructure? Unfortunately, we are not able to answer this question. For example, we do not have consistent information on total train delay minutes that each failure caused, which is an important overall measure of railway performance. A reduction in the number of train delay failures does not *per se* imply that the number of train delay minutes has decreased. Nevertheless, it is fair to say that a reduction in the number of failures causing train delays is a good sign of improved performance (note, however, that the estimate was not statistically significant). Still, the number of failures not causing a train delay has increased quite substantially due to the tilted performance schemes, and the estimated effect was statistically significant. The estimated impact corresponds to 998 more failures per year for the railway network in our sample, while the estimated decrease in the number of failures causing train delays is 102 per year. Possible consequences of this observation need to be further studied, especially since it signals a deteriorating asset condition. For example, what is the impact on the life-cycle cost of the infrastructure when the number of failures (not causing train delays) increase? This should be compared to the impact on delay costs due to the decrease in failures causing train delays. This is especially relevant considering the negative experience in Britain where misaligned incentive structures arguably led to a deteriorating asset condition.

7.0 Conclusion

This paper offers evidence on the effect of different contract designs in rail maintenance services. It contributes to the existing literature by providing empirical evidence on the marginal effect of incentive intensity in the rail maintenance contracts, as well as the effect of tilted performance incentive schemes. Specifically, the results show that a higher incentive intensity (measured as the level when cost reimbursement to contractors applies), reduces the number of failures.

The econometric test of the tilted performance incentive schemes confirms our hypothesis that it influences the relationship between the number of failures causing train delays and other failures. We can conclude that this contract design seems to have been beneficial with respect to the number of train delay failures, yet at the expense of other types of failures which have increased significantly. This increase indicates a deteriorating asset and it can have an impact on the life-cycle cost of the infrastructure; an effect that needs to be evaluated and compared with the benefit of fewer failures causing train delays.

Our findings are informative in considerations on the design of railway maintenance contracts, especially for other IMs across Europe that plan to use (or are already using) competitive tendering. Setting a low incentive intensity can be costly for the IM with respect to the number of failures that occur, while a high incentive intensity can induce rent extraction. This result can be used when making a trade-off between the expected number of failures and the expected maintenance cost of the contract. Moreover, when using tilted performance incentive schemes, the IM needs to contemplate the reallocation of attention from other tasks. For example, its effect on future maintenance costs needs to be estimated and compared with the (short-run) decrease in user (delay) costs. In general, the effect of different designs on cost efficiency in railway maintenance - considering both user and producer costs - is an area for future research. Such considerations are critical in the study of optimal contract design within this field.

Acknowledgements

The author is grateful to the Swedish Transport Administration (Vivianne Karlsson and Anders Nilsson) for providing data. Special thanks to Jan-Eric Nilsson, Svante Mandell, Lars Hultkrantz and Johanna Jussila Hammes who have provided helpful comments on earlier versions of this paper. The paper has also benefited from comments by participants at the First

Nordic Meeting in Transport Economics in Oslo, November 26, 2014, at a seminar at VTI, Stockholm, March 4, 2015, at the ITEA conference on Transportation Economics in Oslo, June 17, 2015, and during the 3rd Meeting on Transport Economics and Infrastructure, Barcelona, January 18, 2019. All remaining errors are the responsibility of the author.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Allison, P. D., Waterman, R. P., 2002. Fixed-Effects Negative Binomial Regression Models. Sociological Methodology. 32, 247-265.

Armstrong, M., Sappington, D. E. M., 2007. Recent developments in the theory of regulation, in: Armstrong, M., Porter, R. H., (Eds.), Handbook of Industrial Organization, Volume 3. North-Holland, Elsevier, pp. 1560-1678.

Banverket, 1997. Spårlägeskontroll och kvalitetsnormer – Central mätvagn Strix. Föreskrift BVF 587.02, (in Swedish).

Banverket, 2000. Utvecklingsstrategi för Banverkets produktionsverksamhet: Förutsättningar och strategier för konkurrensutsättning av Banverkets produktionsverksamhet. GD-stab, GD00-2827/01, 2000-09-25 (In Swedish).

Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics. Methods and applications. Cambridge University Press.

Dalen, D. M., Gómez-Lobo, A., 2003. Yardsticks on the road: Regulatory contracts and cost efficiency in the Norwegian bus industry. Transportation. 30, 371-386.

Devers, C. E., Canella Jr., A. A., Reilly, G. P., Yoder, M. E., 2007. Executive Compensation: A Multidisciplinary Review of Recent Developments. J. of Management. 33(6), 1016-1072.

Domberger, S., Meadowcroft, S., Thompson, D., 1987. The Impact of Competitive Tendering on the Costs of Hospital Domestic Services. Fisc. Stud. 8(4), 39-54.

Gagnepain, P., Ivaldi, M., 2002. Incentive regulatory policies: the case of public transit systems in France. RAND J. of Economics. 33(4), 605-629.

Hausman, J., Hall, B.H., Griliches, Z., 1984. Econometric Models for Count Data with an Application to the Patents-R & D Relationship. Econometrica. 52(4), 909-938.

Hilbe, J.M., 2011. Negative binomial regression. Second edition. Cambridge University Press.

Holmström, B., Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. J. of Law, Economics, & Organization. 7, 24-52.

Laffont, J.-J., Martimort, D., 2002. The Theory of Incentives: The Principal-Agent Model. Princeton University Press, Princeton, New Jersey.

Laffont, J.-J., Tirole, J., 1986. Using Cost Observation to Regulate Firms. J. of Political Economy. 94(3), 614-640.

Laffont, J.-J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. The MIT Press, Cambridge, Massachusetts.

Lazear, E. P., Oyer, P., 2013. Personnel Economics, in: Gibbons, R., Roberts, J. (Eds), Handbook of Organizational Economics. Princeton University Press, Princeton and Oxford, pp. 479-519.

Lewis, G., Bajari, P., 2014. Moral Hazard, Incentive Contracts, and Risk: Evidence from Procurement. Rev. of Economic Stud. 81(3), 1201-1228.

Mundlak, Y., 1978. On the Pooling of Time Series and Cross Section Data. Econometrica. 46(1), 69-85.

OECD, 2017. Government at a glance 2017. OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/gov_glance-2017-en

Odolinski, K., Smith, A. S. J., 2016. Assessing the cost impact of competitive tendering in rail infrastructure maintenance services: evidence from the Swedish reforms (1999-2011). J. of Transp. Economics and Policy. 50(1), 93-112.

Piacenza, M., 2006. Regulatory contracts and cost efficiency: Stochastic frontier evidence from the Italian local public transport. J. of Productivity Analysis. 25, 257-277.

Podgursky, M. J., Springer, M. G., 2007. Teacher performance pay: a review. J. of Policy Analysis and Management. 26(4), 909-949.

Rosenthal, M. B., Frank, R. G., 2006. What Is the Empirical Basis for Paying for Quality in Health Care?. Med. Care Res. and Rev. 63(2), 135-157.

Roy, W., Yvrande-Billon, A., 2007. Ownership, Contractual Practices and Technical Efficiency: The Case of Urban Public Transport in France. J. of Transp. Economics and Policy. 41(2) 257-282.

Salanié, B., 2005. The Economics of Contracts, A Primer, Second edition, The MIT Press, Cambridge, Massachusetts.

Smith, A.S.J., Wheat, P., 2012. Evaluating Alternative Policy Responses to Franchise Failure. Evidence from the passenger rail sector in Britain. J. of Transp. Economics and Policy. 46(1), 25-49.

StataCorp., 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

Stenbeck, T., 2008. Quantifying Effects of Incentives in a Rail Maintenance Performance-Based Contract. J. of Construction Engineering and Management. 134(4), 209-516.

Van Herck, P., De Smedt, D., Annemans, L., Remmen, R., Rosenthal, M. B., Sermeus, W., 2010. Systematic review: Effects, design choices, and context of pay-for-performance in health care. BMC Health Services Res. 10:247.

Vickerman, R., 2004. Maintenance incentives under different infrastructure regimes. Utilities Policy. 12, 315-322.

Wooldridge, J.M., 2013. Introductory Econometrics: A Modern Approach, fifth edition, South-Western, CENGAGE Learning.

Wunsch, P., 1994. Estimating Menus of Linear Contracts for Mass Transit Firms, Mimeo, CORE.

Appendix

Table 4 - Results Model 1: Poisson conditional fixed effects

	IRR	Std. Err.	[95 % Conf.	Interval]
Cost-cap	0.9727**	0.0127	0.9481	0.9978
Rail weight	1.5907***	0.2625	1.1511	2.1982
(Rail weight)^2	0.9911***	0.0031	0.9850	0.9972
Quality class	1.0407	0.0562	0.9361	1.1569
Track length	1.0000	0.0000	1.0000	1.0000
Million gross ton density	1.0479***	0.0143	1.0204	1.0762
(Million gross ton density)^2	0.9982***	0.0006	0.9971	0.9993
D.year2005	1.5147***	0.1883	1.1872	1.9327
D.year2006	1.3949***	0.1792	1.0844	1.7943
D.year2007	1.7151***	0.2206	1.3329	2.2069
D.year2008	1.6101***	0.2063	1.2526	2.0697
D.year2009	1.4999***	0.1924	1.1664	1.9287
D.year2010	1.4805***	0.1893	1.1523	1.9021
D.year2011	1.4497***	0.1873	1.1255	1.8674
D.year2012	1.4464***	0.1886	1.1201	1.8676
D.year2013	1.4821***	0.1956	1.1442	1.9196

***, **, *: Significance at 1%, 5%, 10% level, respectively

Log likelihood: -12 575.699

Number of observations = 7 801

Table 5 - Results Model 2 and 3: Poisson conditional fixed effects

	Model 2 - Train delay failures		*Model 3 - Other failures*	
	IRR	Rob. Std. Err.	IRR	Rob. Std. Err.
D.Mix tendered and not tendered in comp.	1.0414	0.0437	1.0888***	0.0268
D.Year tendered in competition	0.9983	0.0490	1.0871***	0.0281
D.If tendered in comp. in 2003-2013	1.0334	0.5573	0.9826	0.3565
Rail weight	1.4809	0.3710	0.9544	0.1497
(Rail weight)^2	0.9921*	0.0047	1.0005	0.0031
Quality class	1.1216	0.0818	1.0280	0.0444
Track length	1.0000	0.0000	1.0000	0.0000
Million gross ton density	1.0361**	0.0170	1.0232*	0.0134
(Million gross ton density)^2	0.9993	0.0007	0.9990*	0.0006
D.year2004	0.9178**	0.0348	0.8982***	0.0177
D.year2005	0.9318	0.0447	0.8340***	0.0232
D.year2006	0.9346	0.0402	0.8005***	0.0220
D.year2007	1.1136**	0.0557	0.9805	0.0316
D.year2008	1.0264	0.0617	0.9333**	0.0314
D.year2009	0.9478	0.0621	0.9342*	0.0329
D.year2010	1.2041***	0.0739	0.8938***	0.0335
D.year2011	1.2692***	0.0839	0.8580***	0.0329
D.year2012	1.1417**	0.0765	0.7402***	0.0298
D.year2013	1.3246***	0.0882	0.7384***	0.0327

***, **, *: Significance at 1%, 5%, 10% level, respectively

Log likelihood: *Model 2* = -20 561.523, *Model 3* = -42 150.520

Number of observations: 19 561 in *Model 2* and 23 527 *Model 3.*