



K2 WORKING PAPER 2021:7

Evaluating bids on price and quality

The impact on performance in Swedish public bus service

Ivan Ridderstedt and Roger Pyddoke



Date: May 2021
ISBN: 978-91-986324-2-2
Print: Media-Tryck, Lund

Any conclusions and recommendations expressed are the author's own and do not necessarily reflect the views of K2.

K2 WORKING PAPER 2021:7

Evaluating bids on price and quality

The impact on performance in Swedish public bus service

Ivan Ridderstedt and Roger Pyddoke

Table of contents

Preface	5
Abstract.....	7
1. Introduction	9
2. Scoring rules in Swedish procurement of public bus service.....	12
3. Method and data	14
3.1. Empirical framework	14
3.2. Data	16
3.3. Estimation strategy	20
4. Results	22
5. Discussion	25
6. Conclusions	27
7. Acknowledgements	28
8. Declaration of interest	29
References	30
Appendix	33

Preface

This study is the fourth part of the overarching project Incentives and Evaluation for Improved Public Transport. It has been conducted by VTI with funding from and in collaboration with K2 – the Swedish Knowledge Centre for Public Transport. The authors are grateful to Västtrafik for their cooperation and for providing data. The paper also benefited from discussions and comments related to our presentations at the Thredbo 16 conference on competition and ownership in land passenger transport in August 2019, the VTI review seminar in November 2019 with Professor Hans Lööf as the reviewer, and the Transportforum 2020 conference.

Stockholm, May 2021

Roger Pyddoke

Project leader

Abstract

The standard public procurement approach has been to award contracts to the lowest bidder meeting the requirements. In the EU, another award mechanism is almost as common and, moreover, is being promoted, i.e. awarding contracts to the best combination of price and quality, though whether this provides better value has rarely been studied. We contribute relevant empirical evidence by analyzing performance in public bus service contracts. In awarding contracts, a Swedish regional public transport authority alternated between favoring the lowest price and evaluating both price and bus operators' processes for managing and improving bus service. The results of comparing these two approaches suggest that bus operators perform equally well in terms of punctuality regardless of the award mechanism. We conclude that caution is needed when promoting and implementing award mechanisms evaluating both price and quality. Knowledge of whether and how these outperform lowest price mechanisms remain limited in practice.

In Swedish

Standardmetoden för offentlig upphandling har tidigare varit att tilldela kontrakt till den lägsta budgivaren som uppfyller kraven. I EU främjas en annan tilldelningsgrund som redan är nästan lika vanlig, tilldelning av kontrakt till bästa kombination av pris och kvalitet, men om detta ger ett bättre värde har sällan studerats. Vi bidrar med empiriska resultat om hur valet av tilldelningsgrund påverkar utfall i svenska kollektivtrafikavtal. En svensk regional kollektivtrafikmyndighet har alternerat mellan "lägsta pris" och utvärderingsmodeller som täcker både pris och bussoperatörers processer för att hantera och förbättra busstrafiken. Resultaten av att jämföra dessa två tillvägagångssätt tyder på att bussoperatörer presterar lika bra när det gäller punktlighet oavsett tilldelningsgrund. Vi drar slutsatsen att försiktighet krävs när rekommendationer ges om och implementering görs av en tilldelningsgrund som medför utvärdering av både pris och kvalitet. Erfarenhetsbaserad kunskap om huruvida en sådan tilldelningsgrund är överlägsen "lägsta pris" är tills vidare begränsad.

1. Introduction

For citizens around the globe, how their government agencies procure has notable effects on the cost and quality of public services and even on the characteristics of the private sector. With few exceptions, governments combine public ownership and production with substantial procurement of goods and services from the market. In OECD countries, public procurement represents on average 29.1% of total government expenditures (OECD, 2017). Consequently, public procurement policies have a significant impact on welfare.

The mechanism for allocating contracts is decisive in public procurement. Historically, the standard approach has been to award a contract to the lowest bidder meeting a set of requirements. However, many public procurers use another award mechanism: awarding the contract to the bidder offering the best combination of price and one or several dimensions of quality. There are many terms for this alternative approach.¹ We use the term “scoring rule award mechanisms” or just “scoring rules,” referring to the mathematical formula required to derive a rankable score accounting for multiple dimensions.

The aim of this paper is to contribute empirical findings regarding the effectiveness of scoring rule award mechanisms. There is a considerable knowledge gap regarding whether and how these outperform other award mechanisms. The examined case is a Swedish regional public transport authority (PTA) that has alternated between a scoring rule award mechanism and lowest price in procuring public bus service, facilitating a comparative analysis of the two award mechanisms. The PTA’s aim in using scoring rules is to improve several areas of operator performance. Given the breadth of these scoring rules, it is beyond the scope of one study to thoroughly examine all potential quality outcomes.

Our research question is: *How did the use of scoring rules influence the punctuality of the bus service?* The bidders’ plans for achieving reliable bus service are one area evaluated by the PTA. Moreover, punctuality has been identified as a central quality dimension of public transport and a strong determinant of traveler satisfaction (Hensher et al., 2003; Susilo and Cats, 2014; van Lierop et al., 2018; Börjesson and Rubensson, 2019). We estimate the association between chosen award mechanism and operator punctuality using multiple linear regression with a set of control variables for other potentially covarying factors.

Scoring rule award mechanisms are common in many European countries. In 2019, scoring rules were used in 42% of the approximately one million contract notices published in the joint European public procurement database, Tender Electronic Daily (TED, 2019). Fig. 1 shows the level and recent trend in use of this mechanism in 31 European countries. Scoring rules are widely used in the region, clearly dominating in several countries. With the most recent directive on public procurement, Directive 2014/24/EU, the European Union (EU) has recommended that member countries use this award mechanism to attain better value.

¹ First, there is variation in whether this aspect of the contract design is termed “award mechanism,” “award criteria,” “selection criteria,” or “evaluation criteria.” Second, there is variation in whether the mechanism is termed “multi-attribute,” “multi-dimensional,” “scoring rule,” “scoring,” “best value,” or “A + B.” In the EU, it has been called the “economically most advantageous tender” (EMAT or MEAT), which also appears in the academic literature. This term has now been replaced with the term “best price-quality ratio” in the EU terminology (Directive 2014/24/EU).

Moreover, the scoring rule award mechanism is highlighted as an important tool in the EU’s Green Public Procurement initiative (EU, 2016).

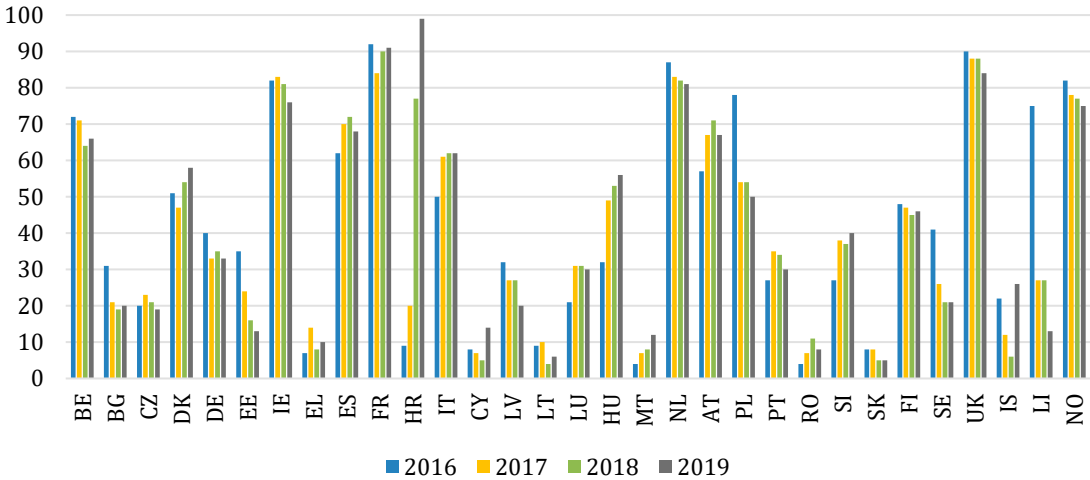


Fig. 1. The share of public procurements awarded using a scoring rule award mechanism. Public procurements listed in the EU’s TED database (TED, 2019)

To our knowledge, there is no database or literature covering the choice of award mechanism in regions other than Europe. However, we have identified several instances indicating that scoring rules are also common outside Europe, in a range of areas. For instance, both the U.S. and Australian departments of defense aim to prioritize value over price by mainly procuring using scoring rules accounting for both price and several qualities in the offers (U.S. Government Accountability Office, 2017; Australian Department of Defense, 2020).

Scoring rules can be formulated to reflect the procurer’s preference for price and one or multiple qualities, thereby accommodating competition in multiple dimensions and contributing to the procurer’s utility. A lowest price auction only accommodates competition in price, with the quality level being determined by the contract specifications and quality requirements. Hence, a scoring rule award mechanism can improve the procurer’s utility and efficiency if bidders are heterogenous in multiple dimensions, i.e. if their competitive advantages lie in different qualities (e.g. Che, 1993; Asker and Cantillon, 2008, 2010). Moreover, the scoring rule award mechanism does not require accurate information about the cost of quality in order to align the price–quality outcome with procurer preferences. The procurer need only formulate a scoring rule reflecting its valuation of quality. In contrast, in a lowest price auction, a procurer with limited information about the cost of quality risks setting the quality requirements either too high or too low (e.g. Bergman and Lundberg, 2013).

Despite the promotion, widespread use, and theoretical merits of scoring rules, there are few empirical studies of whether they outperform other award mechanisms. Previous empirical research into scoring rules has mainly examined their implementation, focusing on the mathematical properties of the scoring rule formula (e.g. Waara and Bröschner, 2006; Bergman and Lundberg, 2013; Stilger et al., 2017; Ridderstedt, 2021). One general finding of this literature is that many scoring rules found in practice are unlikely to reflect procurer preferences. A corollary is that there is large variation in what qualities are evaluated. Even in a scoring rule auction, most qualities are governed by contract specifications and quality requirements. Hence, part of the policy choice when implementing scoring rules is deciding which of many qualities to include in the bid evaluation.

We have identified three empirical studies examining performance-related effects of scoring rules. Lewis and Bajari (2011) studied the effect of scoring rules evaluating price and offered time to completion, which are often used in the U.S. for road maintenance contracts. The examined case is procurements by the California Department of Transportation (DOT), but similar scoring rule variants are used in several DOTs in the U.S. (e.g. New York State DOT, 1999; California DOT, 2002; Minnesota DOT, 2005; Texas DOT, 2018). Contracts tendered using this scoring rule were found to be completed 30–40% faster, with welfare gains to commuters exceeding the increased procurement cost.

Koning and Meerendonk (2014) studied the effects of changes in the scoring weights, i.e. the impacts assigned to various evaluated qualities. Their case is Dutch auctions of workforce reintegration services for the unemployed and disabled with a scoring rule covering three qualities. Results of a regression analysis indicate that increased weights on past performance and the proposed methodology for reintegration were associated with more job placements for workers. The third quality, the stated likely rate of job placement, was not found to have an effect.

Decarolis et al. (2018) studied the effect of a scoring rule implemented by an Italian private multi-utility firm. The scoring rule involved evaluating 136 parameters of suppliers' past performance. Compliance was found to increase from 25% to 80%, with a larger effect on parameters given a higher weight in the scoring rule. The study provides some evidence of higher prices when suppliers had higher past performance scores, but this price effect is significantly lower than the estimated benefit of better performance.

Swedish PTAs have implemented variants of scoring rule award mechanisms for several years. To our knowledge, there is no empirical evidence regarding their effects. However, previous empirical studies have shown that some related policies for improving Swedish public bus service, such as monetary incentives for punctuality and increased ridership, have not performed as intended (e.g. Jansson and Pyddoke, 2010; Pyddoke, 2020; Vigren and Pyddoke, 2020). These findings are in line with experience in other countries of similar efforts to incentivize better performance in public transport (e.g. Batarce and Ávila, 2020).

In public procurement, there are multiple possible reasons why delivered quality could deviate from what is offered and evaluated in the bidding phase. One main factor is that contracts are incomplete. It is prohibitively costly to stipulate the action of each party for every conceivable eventuality (e.g. Hart and Moore, 1988; Holmström and Tirole, 1989). Another important factor is asymmetric information. Typically, a seller has better information about its performance than does the procurer; the seller may then be able to profit from delivering less quality than it offered. A natural remedy for such a moral hazard strategy is for the procurer to limit the information gap through monitoring. However, complete monitoring and penalizing of moral hazard is generally impossible or prohibitively costly (e.g. Holmström, 1979). Adding to the procurer's challenges, an auction with scope for moral hazard behavior risks favoring sellers with a more aggressive moral hazard strategy. This may cause a systematic adverse selection of winning bidders. These general contracting issues emphasize the need to evaluate scoring rule award mechanisms with respect to monitored performance, and not only their conceptual merits or quality offered in the bidding phase.

The remainder of the paper is structured as follows. Section 2 describes Swedish governance of public bus service and the scoring rule award mechanism examined here. The method and data are presented in section 3. In section 4, the results concerning the effect of the scoring rule award mechanism are presented. The results are discussed in section 5. In section 6, we present our conclusions regarding policy implications and areas for further research.

2. Scoring rules in Swedish procurement of public bus service

In Sweden, public bus service is the responsibility of 21 independent government agencies at the regional level, i.e. public transport authorities (PTAs). Since 1989, the PTAs can decide whether and how to use competitive tendering. In 2020, nineteen of the PTAs used competitive tendering, with several PTAs using scoring rule award mechanisms. Ridderstedt (2021) found that about one third of the Swedish public bus service contracts active in 2015 were tendered using some variant of a scoring rule award mechanism rather than lowest price. Moreover, Ridderstedt's (2021) survey found large differences between PTAs in whether and how a scoring rule award mechanism was implemented; however, each PTA was generally consistent in using either lowest price or scoring rule.

This study examines one PTA, Västtrafik, owned by Västra Götaland region (VGR). Västtrafik is the only Swedish PTA that, over a longer period, has alternated evenly between lowest price and variations of a scoring rule award mechanism (Ridderstedt, 2021), facilitating analysis of the present research question. Located in southwest Sweden, with about 1.7 million residents, VGR is Sweden's second most populous region (Statistics Sweden, 2020a). The largest city in the region, Gothenburg, is Sweden's second most populous city with about 580,000 residents (Statistics Sweden, 2020b).

Västtrafik's strategies and management are governed by the goals of VGR (Västtrafik, 2020). VGR's overarching goal for its public transport is for the share of sustainable travel to increase in the region (VGR, 2016). Sustainable travel includes several travel modes, such as public transport, cycling, and carpooling. For public transport, numerical targets are set for both the market share and total number of trips per year. Notably, the latter has been targeted to double between 2006 and 2025. VGR has a set of sub-goals for how to work towards the overarching goal. Between 2016 and 2020, these sub-goals concerned accessibility, public transport attractiveness, inclusion, and environmental impact. Two groups of factors have been prioritized for improving public transport attractiveness: the first concerns how smart, safe, and convenient travelers perceive sustainable travel to be; the second concerns service punctuality and reliability (VGR, 2016).

Västtrafik's use of scoring rule award mechanisms is part of a larger procurement strategy that has evolved gradually since these mechanisms were introduced in 2011. This strategy emphasizes close cooperation with partners, trust, passenger incentives, and procurement of function rather than detailed specifications (Västtrafik, 2011). Västtrafik (2020b) has stated that the steering group for each procurement decides the award mechanism and that scoring rules are generally used for traffic assignments involving passenger incentives. Accordingly, such incentives are mainly used for traffic assignments in which Västtrafik has identified good potential for increasing ridership. As stated in the following translated quotation about Västtrafik's procurement strategy, the aim of using scoring rules is to award contracts to the best partner:

The procurer's strategy is to continue to move from traditional procurer and supplier roles to partnership for joint business development. One part of this effort is to use a new model of quality evaluation, intended to select the "best" partner, in new tendering. The model is also a tool for achieving Västtrafik's overall goals and the parties' joint goals for the traffic commitment. (Västtrafik, 2015, p. 3 in Appendix A2)

With a similar degree of generality, Västtrafik also stated that the aim of using scoring rules is to select the most suitable operator for the assignment throughout the contract and to incentivize continuous improvement (Västtrafik, 2013).

Västtrafik's broad aim of using scoring rules is reflected by the number and breadth of evaluated qualities. The scoring rules implemented by Västtrafik are based on a cross-sectoral model of customer-oriented business development, the SIQ model, widely adopted in Sweden since the 1990s (Swedish Institute for Quality, 2017). This model highlights a broad range of business culture values and the need for structured processes for management, evaluation, and improvement. Correspondingly, with Västtrafik's SIQ-based scoring rules, operators are required to describe multiple aspects of how they will fulfil the contract. The operators do this by answering 13–22 questions arranged in thematic categories (e.g. Västtrafik, 2011, 2013, 2015). For contracts with a traffic start in 2016, the categories were:

- Establishing traffic start
- Control and management
- Traffic and market development
- Preparing traffic delivery
- Performing traffic delivery

Similar categories were used in the prior procurement. However, some procurements have included additional categories with questions about how to increase the number of passengers and their satisfaction, how to improve cost-effectiveness, and previous results (e.g. Västtrafik, 2011, 2013).

Compared with the scoring rules used by other Swedish PTAs, Västtrafik's implementations are among the most complex (Ridderstedt, 2021). For instance, each of the 13–22 answers are graded in two dimensions, initially *Approach* and *Evaluation and improvement* and more recently *Extent* and *Level* (Västtrafik, 2011, 2015). Consequently, 26–44 individual grades are set for each offer. For several years, Västtrafik assigned varying weights to both each question and the two dimensions of grading; however, in more recent procurements, all qualities have been assigned the same weight (Ridderstedt, 2021).

The evaluation is conducted by an evaluation team comprising both internal and external reviewers (e.g. Västtrafik, 2011, 2015). In more recent tendering documents, Västtrafik has specified that the operators' submitted texts should first be assessed individually and finally in a consensus meeting (Västtrafik, 2015). The evaluated quality of the winning bid, or a negotiated version of it, is contractually binding (Västtrafik, 2015). The contracts specify that Västtrafik will conduct at least one on-site control of whether and to what extent the operator is implementing the processes as offered. The penalty for obtaining a considerably worse than agreed-on quality score in the control evaluation is EUR 100,000–1,000,000, depending on the magnitude of the deviation and on the contract. If such deviations are severe, Västtrafik retains the right to terminate the contract (e.g. Västtrafik, 2011, 2015).

3. Method and data

3.1. Empirical framework

In contrast to the main theoretical rationale for scoring rules, Västtrafik's aim is primarily to improve quality rather than efficiency or value for money. Correspondingly, the aim of this paper is to contribute empirical findings regarding the effectiveness of scoring rule award mechanisms in terms of improving quality rather than efficiency. However, given the breadth of qualities covered by Västtrafik's scoring rules, as described in section 2, it is beyond the scope of one study to thoroughly examine all the potential quality effects; instead, this study focuses on one area of operator performance.

The number of passengers and the increase thereof are performance outcomes closely related to Västtrafik's overarching goals. However, in reviewing empirical evidence on what explains public transport ridership, Taylor and Fink (2013) concluded that ridership is mainly affected by factors external to the public transport manager, and thus to the operator as well. In particular, these external factors are car ownership and factors affecting it (e.g. Kitamura, 1989; Liu, 1993; Kain and Liu, 1996), such as parking availability and pricing (e.g. Morral and Bloger, 1996; Shoup, 2005; Guo, 2013). Other external factors highlighted by Taylor and Fink (2013) are factors such as population density and growth, employment levels, and household income. With many other factors strongly affecting ridership, it would be difficult and require exceptionally rich data to isolate the association between Västtrafik's scoring rules and ridership. The argument for not making customer satisfaction the examined performance outcome is similar: it is an outcome dependent on multiple factors, on some of which the operator has only limited influence. For instance, Börjesson and Rubensson (2019) found that crowding and service frequency are two important determinants of customer satisfaction.

Instead, this study focuses on punctuality. In public transport, this and other closely related reliability outcomes have been found to be among the most, or even the most, important determinants of customer satisfaction (e.g. Börjesson and Rubensson, 2019). As described in section 2, these factors are also highlighted as important for public transport attractiveness in the main steering document for public transport in Västra Götaland region. While punctuality is affected by many other factors, it is closely linked to operator performance. For instance, the operator's maintenance and coordination of buses, driver rostering practice, and managerial skills and resources to adapt to sudden changes in preconditions all affect the probability and severity of delays and canceled departures. Correspondingly, Västtrafik's scoring rules involve questions about how the operator will secure an adequate vehicle supply throughout the contract. Moreover, in some of Västtrafik's procurements, the scoring rules involved questions explicitly about how the operator would provide reliable and punctual traffic (e.g. Västtrafik, 2015).

A drawback of measuring delivered quality as punctuality is its dependence on the timetables, which are set and adjusted yearly by the PTA. This characteristic is not shared with the related travel time variability measure, which captures the variation in travel time under similar conditions, in terms of route, day, hour, or even vehicle, and would likely have been a good alternative for this study. However, Västtrafik could not provide this information. Hence, timetable robustness is a factor to account for when comparing punctuality between different traffic assignments.

The hypothesis tested here is that punctuality was higher in traffic assignments tendered using a scoring rule instead of a lowest price award mechanism, all else equal. Our empirical framework for testing this hypothesis can be formulated as follows:

$$PUNCT_{it} = \beta_0 + \beta_1 SRULE_i + \omega' c_i + \theta' f_{it} + \mu_i + \varepsilon_{it}, \quad (1)$$

where $PUNCT_{it}$ is the punctuality of the bus operator in traffic assignment i in month t . β_1 is the association between the use of a scoring rule award mechanism ($SRULE = 1$) and the punctuality in the traffic assignment; c and f are vectors of variables capturing other contract characteristics and exogenous factors, respectively, that affect performance, ω and θ are vectors of coefficient parameters, μ_i is the unobserved time-invariant individual effect of the traffic assignment, and ε is the error term. We test the hypothesis, i.e. if β_1 is larger than zero, using the conventional 95% confidence level.

Our priority is to have control variables for contract characteristics and factors that may correlate with the use of a scoring rule award mechanism and punctuality; this is to limit omitted variable bias on β_1 . Another source of bias that is more difficult to assess and mitigate is whether Västtrafik has systematically selected scoring rule over lowest price procurements for traffic assignments with worse punctuality. As described in section 2, Västtrafik stated that scoring rules are primarily used for traffic assignments involving passenger incentives. The use of these incentives was, in turn, stated to be governed by Västtrafik's assessment of the potential for increasing ridership in the traffic assignment. According to this description, the selection process for implementing scoring rules is not linked to punctuality. We therefore consider that the issue of selection bias is limited.

Previous studies of determinants of punctuality and travel time variability in public bus service offer guidance on what factors to account for in our analysis. Route length, service frequency, time between buses (i.e. headway), and provision of bus lanes have long been established as important internal factors affecting reliability (e.g. Sterman and Schofer, 1976; Abkowitz and Engelstein, 1983, 1984; Strathman et al., 1999). Several studies of similar scope have followed since. We highlight two more recent studies using linear regression analysis. Diab and El-Geneidy (2013) studied determinants of public bus travel time and punctuality using data on 2,250,000 individual stops of buses along two routes in Montreal, Canada, recorded between 2007 and 2011. The number of stops and the amount of rain and snowfall were found to decrease punctuality, whereas reserved bus lanes were found to increase it. Some counter-intuitive results were also found: punctuality was higher in peak hours than mid-day, and the introduction of smart-card ticketing systems and articulated buses decreased punctuality. The authors noted that traffic planners were likely too optimistic about the time savings arising from some new policies and mid-day traffic. Mazloumi et al. (2010) investigated day-to-day travel time variability along a 27-kilometer circumferential bus route in Melbourne, Australia. Weekday data in 2007, a total of 3351 trips, were analyzed. The results indicate that travel time variability was increased by route length, number of traffic signals, number of bus stops, and precipitation; in areas dominated by industry, variability was lower. These results were found during both peak and off-peak hours.

3.2. Data

We contacted Västtrafik with a request for punctuality data for their 72 contracts active in December 2015. These contracts were identified using a data set (Vigren, 2018a) previously analyzed by Vigren (2018b, 2020) and Ridderstedt (2021). We learned that punctuality data were unavailable for most contracts and not recorded at the contract level. Instead, we obtained monthly punctuality data on 19 traffic assignments in and near the city of Gothenburg. Of these traffic assignments, 13 were for express lines and six were for larger city service areas. One procured contract can cover multiple traffic assignments.

Västtrafik defines punctuality as the bus departing from the bus stop less than 30 seconds early and less than 3 minutes late, relative to the timetable. This is monitored at about one third of bus stops (Västtrafik, 2020), i.e. one route departure contributes multiple punctuality observations. Monthly punctuality is the share of departures from these monitored bus stops that were punctual. The data provided by Västtrafik also contain information on the share of departures from the monitored bus stops with successful registration of punctuality, which helps to assess the reliability of the observations.

We obtained monthly observations from January 2010 to April 2019 for the city service areas and from June 2011 until April 2019 for the express lines. We were able to match 18 of the traffic assignments to 15 contracts in our dataset (Vigren, 2018a) and to the related tendering documents, i.e. some traffic assignments were tendered jointly as part of one contract. One traffic assignment (Rosa Express) could not be matched and was therefore excluded.

Table 4 presents an overview of the 15 contracts covering the traffic assignments in our sample. The 18 traffic assignments constitute 84% of the vehicle kilometers in these contracts. Västtrafik's approach of alternating between lowest price and scoring rule award mechanisms is captured in this sub-sample of their contracts. This is a prerequisite for examining our research question. In contrast to Västtrafik's claim of typically using scoring rules in combination with passenger incentive payments, all but two of the contracts tendered using the lowest price mechanism also incorporated these incentives. This is beneficial as it allows us to estimate the effect of the scoring rules on punctuality without capturing the effect of the passenger incentives.

There is considerable variation in the yearly contract value, number of lines, and planned yearly vehicle kilometers among both lowest price and scoring rule contracts. Contracts tendered using a scoring rule award mechanism were generally slightly larger in terms of production and contract value, in line with Västtrafik's claim that it uses scoring rules for more complex traffic assignments. As such complexity may be associated with punctuality, these are factors to control for in the analysis. Even though Västtrafik's implementation of scoring rules involves evaluating qualities that are not correspondingly captured by the minimum requirements in their lowest price tendering, the mean contract cost per kilometer is about the same in both award mechanisms.

Table 4. Contracts represented in the analysis; exchange rate: EUR 1 = USD 1.13 = SEK 10.66 (2019-06-11).

ID	Traffic start	Scoring rule	Passenger incentives	Winning bid per year, EUR millions	Lines in contract	Vehicle km. per year, millions.	EUR per vehicle km
1	2010	1	1	31	32	8	4.0
2	2010	1	1	6	12	4	2.7
3	2010	1	1	8	6	3	2.4
4	2010	1	1	2	1	1	2.7
5	2011	0	1	8	15	5	1.6
6	2011	0	1	6	1	3	2.3
7	2012	0	1	9	17	3	3.3
8	2012	0	1	12	6	3	4.7
9	2013	1	1	32	23	9	3.6
10	2013	1	1	6	1	2	3.2
11	2014	1	1	2	1	1	1.9
12	2014	0	0	4	4	2	2.4
13	2015	0	0	3	9	1	2.3
14	2016	1	1	8	2	4	2.1
15	2016	1	1	5	2	3	2.1
Mean: Lowest price			0.67	7	9	3	2.8
Mean: Scoring rule			1.00	12	11	4	2.7

The monthly punctuality delivered in the traffic assignments from July 2016 to December 2018, when all traffic was running in all assignments, is presented in Fig. 2, which shows that there is enough variation in punctuality, both within and between contracts, to study statistically. There is a cyclical pattern within the years, with punctuality generally being better in July and worse in the autumn. We conclude that no traffic assignment or monthly observation of punctuality is an outlier to exclude from the analysis. Punctuality was successfully logged for an average of 98% (and a minimum of 85%) of monthly departures from the monitored bus stops. The punctuality measures presented in Fig. 2 are based on about 7.1 million observations of departures. We found this sampling process to be more than adequate for reliable measurements of monthly punctuality, even though unsuccessful logging of punctuality may be associated with punctuality.

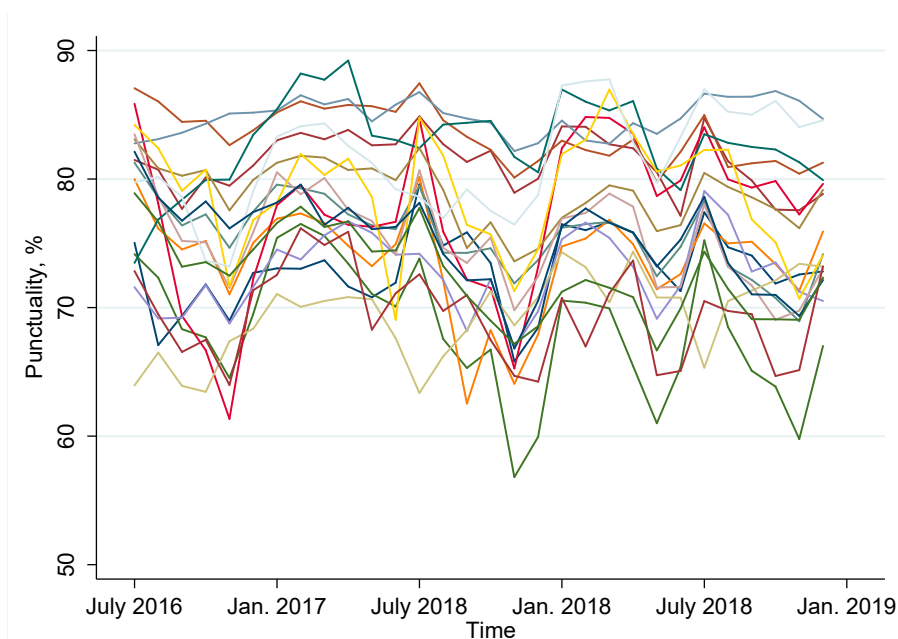


Fig. 2. Monitored monthly punctuality per traffic assignment.

Our statistical analysis is confined to the time period when all traffic assignments were being delivered, i.e. from July 2016 until the end of 2018 when we obtained the data. Observations from earlier contract periods were not included because significant changes in the number of passengers can be observed during the transition. This indicates that the traffic assignments had fundamentally changed in terms of, for instance, routing and service frequency, even though they might serve the same area as previously. These changes between contract periods also limit us from using shifts in the choice of award mechanism for one traffic assignment to estimate the effect of scoring rules.

The descriptive statistics of the variables compiled for the 18 traffic assignments are presented in Table 5. The monthly punctuality, PUNCT, is on average 76%, based on 30 months of observations for each traffic assignment. The indicator variable SRULE captures whether a scoring rule award mechanism (1) or a lowest price award mechanism (0) was used. This information was obtained from the contract dataset (Vigren, 2018a). The five traffic assignments tendered in 2010 included a scoring rule with a non-monetary relative scoring of price, while the seven other implementations of scoring rules were based on quality penalties. We do not distinguish between the two variants in the analysis, as the qualities covered are similar and it would be difficult to isolate the effect of this change in practice from other time-related factors.

The sample covers a diverse set of traffic assignments. Whether the traffic assignment was for a city service area or an express route is captured by the indicator variable CITY, following the definition used by Västtrafik in their data on punctuality. Additional variables capturing important characteristics of the traffic assignments are found in the tendering documents. There is a considerable spread in the number of routes, ROUTES, the average number of bus stops per route, STOPS, and the yearly number of vehicle kilometers, VKM. For one traffic assignment, the vehicle kilometers were not specified in the tendering documents but were found in another Västtrafik document (Västtrafik, 2014).

Västtrafik also provided data on the monthly number of boarding passengers, PASS, per traffic assignment, counted as validated smart-card tickets. On average, the monthly number of boarding passengers is 600,000 but the median is only 250,000, the distribution of this variable being heavily skewed. Fig. 3 plots the monthly number of boarding passengers for the studied period. For a better visual representation, all values are indexed relative to the values for July 2016. As with punctuality, recurring cycles are apparent within the years. For all but one contract, the number of passengers decreases in the summer, typically with the downward peak in July; the upward peaks are in the spring and autumn. The exception is Marstrand Express, which serves the route between Gothenburg and the popular summer destination of Marstrand.

All traffic assignments had penalties related to delays that could be linked to the operator. Two variants were used: one with a EUR 5–10 penalty for each minute of delay (DPEN = 1), and another with a fixed fee of about EUR 125–190 per departure for departures not delivered to at least 90% of the route with at most a 20-minute delay (DPEN = 0, EUR 1 = SEK 10.66). The latter delay penalty variant is less strict about shorter delays. To account for the difference in the operators' experience of delivering traffic on the routes, we add an experience variable, EXPER, capturing the number of previous months of traffic experience as of June 2016.

Lastly, we collected data on the weather, identified in previous studies as an important precondition for punctual bus traffic. Information on monthly mean temperatures, TEMP, and precipitation, PRECIP, observed at a weather station in central Gothenburg was collected from the Swedish Meteorological and Hydrological Institute (SMHI, 2019a, 2019b). The monthly mean temperatures were within the interval of -1.5 and $+21.4$ degrees Celsius. In an average month, the mean daily precipitation, including rain and snow, was 0.5–4.5 millimeters. The

weather variables are mainly included on their own merit, as they are expected to be strong determinants of punctuality that merit quantification. As there is no variation in the weather variables between the traffic assignments, they only affect the estimated effect of scoring rules if the combined effect is analyzed, through the inclusion of an interaction variable.

Table 5 Descriptive statistics.

Variable	Abbrev.	<i>n</i>	Median	Mean	SD	Min.	Max.
Punctuality, %	PUNCT	540	76	76	6	57	89
Passengers, millions	PASSm	540	0.25	0.60	0.89	0.01	3.49
Scoring rule	SRULE	18	1	0.67	0.49	0	1
City service area	CITY	18	1	0.33	0.49	0	1
Routes	ROUTES	18	1	7	10	1	32
Bus stops per route, mean	STOPS	18	27	28	13	5	57
Annual vehicle km, millions	VKM	18	1.9	2.4	2.3	0.5	8.7
Punctuality penalty per minute	DPEN	18	0	0.44	1	0	1
Months active as of June 2016	EXPER	18	67	59	36	1	97
Temperature, °C, mean	TEMP	30	9.3	9.9	6.9	-1.5	21.4
Daily precipitation, mm, mean	PRECIP	30	2.2	2.4	1.1	0.5	4.5

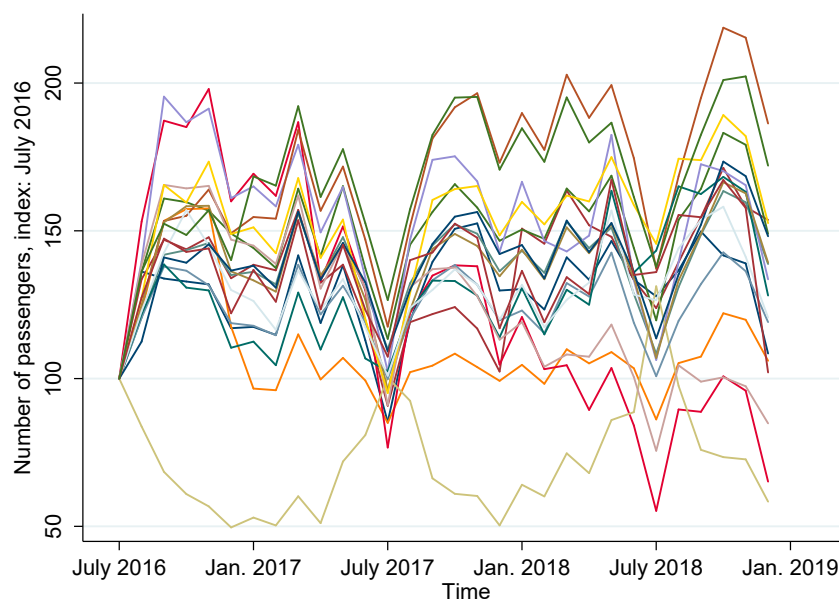


Fig. 3. Monthly number of boarding passengers per traffic assignment.

The correlation coefficients between the variables are presented in Table 6. For variables varying over time, i.e. punctuality and the number of boarding passengers, we use the mean values of the July 2016–December 2018 period. Moreover, the second of these two variables is transformed by the natural logarithm due to its heavily skewed distribution. The weather variables are not included as they only vary over time. Punctuality is most strongly associated with whether the traffic assignment was a city service area or an express route, number of routes, type of delay penalty, and number of stops. The correlation coefficients for scoring rules are relatively low, with the highest being 0.24 (correlation with use of the stricter delay penalty). Still, this could be a source of estimation bias if not controlled for. While several of the variables capturing the characteristics of the traffic assignments are highly correlated, we consider that they still may contribute as separate control variables.

Table 6 Correlation matrix.

	Punkt	In Pass	SRule	City	Routes	Stops	VKM	DPen
In (Pass)	0.14	1.00						
SRule	-0.03	0.10	1.00					
City	0.45	0.80	0.00	1.00				
Routes	0.35	0.69	0.17	0.85	1.00			
Stops	-0.38	-0.46	0.09	-0.44	-0.39	1.00		
VKM	-0.04	0.78	0.16	0.67	0.67	-0.37	1.00	
DPen	0.45	0.72	0.24	0.71	0.60	-0.57	0.48	1.00
Exper	-0.06	0.66	-0.12	0.48	0.42	-0.43	0.31	0.40

3.3. Estimation strategy

We use multivariate linear regression analysis to estimate the association between punctuality and the use of a scoring rule award mechanism. It could be questioned whether it is adequate to model the punctuality ratio with a linear model, in which the predictions can fall outside the possible range. However, the punctuality observations are rather narrowly distributed around a mean of 76%, and no observations are very close to either 0 or 100, so we expect that a linear model will be adequate.

The variable for the use of a scoring rule and all but one of the control variables are time-invariant at the contract level. Hence, we first treat the data as cross-sectional with one observation per traffic assignment, using mean values for the time-variant variables, i.e. punctuality and the number of boarding passengers. As the weather variables only vary on the time dimension, they are not included. This results in the following empirical model:

$$PUNCT_i = \beta_0 + \beta_1 SRULE_i + \beta_2 \overline{\ln(PASS)}_i + \beta_3 VKM_i + \beta_4 CITY_i + \beta_5 ROUTES_i + \beta_6 STOPS_i + \beta_7 DPEN_i + \beta_8 EXPER_i + e_i \quad (2)$$

where \overline{PUNCT}_i and $\overline{\ln(PASS)}_i$ are the mean values of the variables for traffic assignment i during the observed period. β_j is the estimated intercept and the association between punctuality the j explanatory variables, and e_i is a residual term. This model, and a reduced variant with only the use of a scoring rule as an explanatory variable, is estimated using OLS.

The data can also be treated as a panel with 30 monthly punctuality observations for each traffic assignment. In this way, an unobserved time-invariant individual effect can be included in the model. This individual effect accounts for systematic differences in punctuality between the traffic assignments due to time-constant factors for which we lack data. Such factors could be the existence of bus lanes, traffic density, the number of road junctions, etc. This individual effect might also control for some of the differences in the timetabling, although these can be adjusted between the years. The panel structure also allows us to add variables for the monthly number of boarding passengers per traffic assignment, year, and month, and for weather in a more elaborate model of punctuality:

$$PUNCT_{it} = \beta_0 + \beta_1 SRULE_i + \beta_2 \ln(PASS_{it}) + \beta_3 VKM_i + \beta_4 CITY_i + \beta_5 ROUTES_i + \beta_6 STOPS_i + \beta_7 DPEN_i + \beta_8 EXPER_i + \beta_9 TEMP_t + \beta_{10} PRECIP_t + \beta_{11}(TEMP_t * PRECIP_t) + \gamma' YEAR_t + \varphi' MONTH_t + \mu_i + \varepsilon_{it} \quad (3)$$

Time is counted in months starting from July 2016, when $t = 1$. YEAR and MONTH are vectors of dummy variables for all but one of the years and all but one of the months, respectively, and γ and φ are vectors of coefficient estimates; μ_i is the individual effect of traffic assignment i ,

and ε_i is the idiosyncratic error term of i at time t . Only one interaction variable is included, capturing the combined effect of temperature and precipitation, for instance snow.

Model 3 is estimated using pooled OLS, random effects, fixed effects, and correlated random effects. With Pooled OLS, the panel is analyzed as cross-sectional data, without accounting for the individual effects, μ_i . We assess whether the omitted individual effects impose a bias on the estimates using Breusch and Pagan's (1980) Lagrange multiplier test. The random effect estimator is consistent if the explanatory variables are uncorrelated with the individual specific effects. To test whether this is true, we compare the estimates with the estimates of the consistent, but less efficient, fixed effect estimator. In our case, the standard test for deciding between random and fixed effects, the Hausman specification test (Hausman, 1978), is unreliable as it reports that the variance-covariance matrix of random and fixed effects estimates is not positive definite. Hence, we apply a corresponding test based on an artificial regression approach described by Arellano (1993) and Wooldridge (2002, pp. 290–291), using the Stata command `xtoverid` (Schaffer and Stillman, 2010).

With the fixed effect estimator, systematic unobserved differences between contracts are accounted for by subtracting the individual mean from each variable. While this is an effective remedy for this possible endogeneity problem, the de-meaning process cancels all variables that are constant over time at the individual level. A given feature of our data is that the variable of interest, i.e. whether a contract was tendered using a scoring rule or the lowest price award mechanism, does not vary over the time period of any contract. Hence, if the tests suggest that the individual effect is correlated with the explanatory variables, we will apply the correlated random effect estimator, which can be described as a mix of random and fixed effects. Correlated random effects, first proposed by Mundlak (1978), allow for within-estimation (fixed effects) for the time-variant variables and between-estimation (random effects) for the time-invariant variables.

We have also considered applying the estimator of Hausman and Taylor (1981), which can account for a set of both time-variant and -invariant variables being associated with individual effects. The Hausman Taylor estimator is a transformed RE model with instrumental variables to account for the endogeneity problem violating the assumptions of the RE estimator (Hausman and Taylor, 1981; Baltagi et al., 2003; Schunck, 2013). The individual means of exogenous time-variant variables are used as instrumental variables for the RE-transformed time-invariant variables. However, the only time-variant variable that varies between the traffic assignments, i.e. the number of boarding passengers, proved not to be a valid instrument for the scoring rule regressor: first, it could not be rejected that there was no correlation between the two variables; second, overidentification tests for correlation between the instrumental and error terms could not be applied, as they require more instrumental than endogenous regressors.

4. Results

In Table 6, we present the results of six regressions on the monthly punctuality of service, taking a value in the closed interval between zero and one hundred. As a baseline, the results in columns 1 and 2 are from using the OLS estimator for a cross-sectional analysis of the mean punctuality during the July 2016–December 2018 period. In column 3, the results are from applying OLS to 30 monthly observations per contract, i.e. pooled OLS. This model also includes the weather variables and dummy variables for month and year. In columns 4–6, the panel structure of the observations is accounted for with random effects (RE), fixed effects (FE), and correlated random effects (CRE).

Table 7. Results table.

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS ^m	OLS ^m	P. OLS	RE	FE	CRE
SRULE	-0.29 (-0.11)	-0.23 (-0.09)	-0.20 (-0.09)	0.22 (0.07)		-0.23 (-0.10)
ln(PASS)		0.62 (0.32)	0.18 (0.12)	-5.16*** (-5.58)	-6.41*** (-5.99)	-6.41*** (-5.95)
VKM		-1.77* (-1.94)	-1.62*** (-3.14)	0.18 (0.44)		-1.77*** (-3.19)
CITY		6.64 (1.24)	7.06*** (3.01)	12.16*** (3.38)		6.64** (2.50)
ROUTES		0.12 (0.56)	0.11* (1.84)	-0.04 (-0.29)		0.12* (1.81)
STOPS		-0.15 (-1.43)	-0.14 (-1.58)	-0.07 (-0.76)		-0.15 (-1.62)
DPEN		1.01 (0.27)	1.42 (0.58)	6.31*** (2.90)		1.01 (0.40)
EXPER		-0.08 (-1.67)	-0.07** (-2.93)	0.02 (0.50)		-0.08*** (-3.07)
TEMP			-0.10* (-1.97)	-0.02 (-0.30)		0.00 (0.07)
PRECIP			-1.08*** (-3.58)	-1.05*** (-3.50)	-1.04*** (-3.50)	-1.04*** (-3.48)
TEMP* PRECIP			0.09*** (3.15)	0.08*** (3.02)	0.08*** (3.00)	0.08*** (2.98)
Constant	76.58*** (36.00)	86.88*** (9.07)	88.53*** (14.88)	64.60*** (11.37)	68.95*** (34.95)	89.43*** (14.40)
Month	No	No	Yes***	Yes***	Yes***	Yes***
Year	No	No	Yes	Yes	Yes	Yes
<i>N</i>	18	18	540	540	540	540
<i>R</i> ²	Standard 0.00	Standard 0.65	Standard 0.56	Between 0.41	Within 0.44	Between 0.65

Dependent variable: Monthly punctuality $\in [0, 100]$ (i.e. share of monitored departures from bus stops less than 30 seconds early and 3 minutes late). ^m With mean values for PUNCT and PASS. Cluster robust *t* statistics in parentheses. Clusters (15): Traffic assignments tendered as part of the same contract. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The results presented in Table 7 do not offer any evidence for punctuality being different in traffic assignments tendered using a scoring rule award mechanism instead of lowest price. None of the estimates of the scoring rule parameter is statistically significantly different from zero at the 95% confidence level. Moreover, the estimates are generally slightly negative and no estimate is close to being statistically significant.

Still, we apply the tests described in section 3.3 to assess which model and estimator provides the most accurate and consistent estimates. According to the Lagrange multiplier test, we reject the null hypothesis that the individual specific variance components are zero at the 0.001 level, i.e. heterogeneity across traffic assignments is not accounted for adequately with OLS. In the test corresponding to the Hausman specification test, the null hypothesis of no systematic difference between the estimates is rejected at the 0.001 level, favoring fixed effects over random effects. Hence, as described in section 3.3, we find that correlated random effects (column 6) provide the most accurate and consistent estimates.

As a sensitivity analysis, we investigate whether scoring rules had an effect in combination with other contract characteristics and the weather variables. Table A1 in the Appendix presents the results of adding interaction variables between SRULE and each of the control variables, one at a time. This is estimated using CRE. When estimating interaction effects with few observations at the individual level, the estimate is based on a few observations for each combination of the interacted variables. Especially when interacting SRULE with the other dummy variables, individual observations have a large impact on the estimation. Testing all these interactions without a clear theoretical rationale also increases the risk of finding spurious associations within the sample. Still, combined with the results of future studies, these results may still be of value.

Overall, we find that the results of this sensitivity analysis are in line with the main results presented in Table 7. Even at the 90% confidence level, only two interaction effects were statistically significant. The combined effect of the use of scoring rules and the number of months a traffic assignment was active up to July 2016 is statistically significant at the 95% confidence level. According to the results, visualized in Fig. A3 in the Appendix, punctuality was higher in traffic assignments tendered more recently, regardless of whether the lowest price or scoring rule was used. However, of the more recently tendered traffic assignments, punctuality was higher when tendered using the lowest price mechanism. On the other hand, the estimates also suggest that for traffic assignments that started 90 or more months before July 2016, punctuality was higher when tendered using a scoring rule.

The estimated interaction effect of using a scoring rule and monthly precipitation is statistically significant at the 90% confidence level, but not the interaction with temperature or the interaction with temperature and precipitation capturing snow. These results imply that traffic assignments tendered using a scoring rule had worse punctuality in months with high precipitation, relative to those tendered using the lowest price mechanism, but slightly better punctuality in months with low precipitation. This is visualized in Fig. A4, in which temperature is fixed at zero degrees Celsius.

While the empirical strategy focuses on the scoring rule parameter estimate, we also obtain other estimates of interest in relation to the punctuality of bus service. According to the results using the CRE estimator, the association between the number of passengers and punctuality is statistically significant even at the 99% confidence level. The point estimate implies that a one percent increase in the number of passengers is associated with about 6.41-percentage-point lower monthly punctuality, all else equal. The difference between the city service areas and the express lines is also statistically significant, at the 95% level. Punctuality is about 6.64 percentage points higher in the contracts for city service areas than for the express routes. The magnitudes of these estimates are notably large considering that, for instance, the vehicle kilometers and number of routes are controlled for. The association between the vehicle kilometers supplied, relative to both service frequency and route length, and punctuality is also statistically significant, at the 99% level. One million additional yearly vehicle kilometers would, according to the point estimate, be associated with 1.77-percentage-point lower monthly punctuality.

A further finding is that the number of months a traffic assignment was active up to and including July 2016 is negatively associated with punctuality, an association statistically significant at the 99% level. Twelve months of active traffic assignment up to and including the studied period is associated with about 1-percentage-point lower punctuality in the studied period. This is contrary to our expectation of a run-in period for operators when they begin service in a new area. This could be interpreted as indicating that Västtrafik had become better at promoting good performance in later contracts, or that operators' performance deteriorates towards the end of contracts.

The estimate of the number of routes is positive and statistically significant at the 90% level. The estimates of the other two contract parameters are not significantly different from zero in either model, although the signs are as expected. The estimate of the use of minute-based penalties for deviations in punctuality is positive and the estimate of the number of stops is negative. Finally, the results suggest that the weather has a significant influence on punctuality. The amount of precipitation is negatively associated with punctuality, and more so the lower the temperature. No year fixed effects are statistically significant, but several of the month fixed effects are at the 99% confidence level.

We consider that the model specifications perform well in explaining the observed punctuality. With the estimates obtained using the CRE estimator, we obtain within-sample predictions of punctuality in the range of 64–88%, and neither model yields within-sample predictions outside the possible range between zero and one. Using OLS and CRE, the model explains about 65% of the between-traffic-assignment variation in punctuality. Using FE and CRE, about 44% of the within-traffic-assignment variation in punctuality is explained. We consider that the models adequately control for important determinants of punctuality that, if omitted and correlated with the use of scoring rules, would have biased our estimate of interest.

The distributions of the residuals of OLS with control variables, column 2, and the idiosyncratic errors from using the CRE estimator are plotted in Figs. A1 and A2 in the Appendix. Based on Fig. A1, we consider that the OLS model performs similarly well regardless of the traffic assignments' level of punctuality, and that there is no notable problem with the specifications. Fig. A2 indicates that our model also performs well using CRE and the panel data. For the 540 studied observations, the punctuality predictions are typically less than 5% off. There are some deviant overpredictions, differing by over five or even ten percent from the observations. These are distributed over 13 different traffic assignments, i.e. the most severe overpredictions are likely due to stochastic events rather than a misspecification of the model. Fig. A2 also shows that the ten highest predictions of punctuality are all overpredictions compared with the observations, which is not ideal. Still, we find that these issues have a minor impact on the estimate of the scoring rule parameter.

5. Discussion

Västtrafik's implementation of scoring rules was intended to select the best partner, understood as an operator providing better performance in several aspects of the bus service. This broad aim and the evaluation of intermediary processes rather than observable user-end outcomes make it challenging to assess the effectiveness of the policy. In this study we have focused on punctuality, previously identified as an important quality aspect affecting customer satisfaction with public transport. Moreover, Västtrafik's scoring rules involve evaluation of the operators' offered processes about or closely related to punctuality. According to the present results, punctuality was no better in the contracts procured using a scoring rule award mechanism instead of the lowest price. While the results do not exclude the possibility that the sampled contracts delivered improved quality in other dimensions, our inference from the results is that the PTA's scoring rule award mechanism did not outperform the lowest price mechanism. This is contrary to previous empirical findings regarding the effects of other implementations of scoring rules.

In the analyzed sample, the scoring rule award mechanisms covered qualities in addition to lowest price. Furthermore, the winning bidders typically obtained high quality scores, well above the minimum level. Thus, there appears to be a gap between the evaluated quality and delivered quality, the reason for which is unclear. One possible explanation is that the evaluated processes and how they were graded do not capture important factors determining good performance, at least not in terms of punctuality. Another possible explanation is that the operators delivered processes inferior to those offered, i.e. moral hazard. Verifying and incentivizing quality throughout the contract is a well-acknowledged issue in contracting in general. However, Västtrafik has attempted to deter such problems by conducting a control review and having the right to conduct additional control reviews, impose penalties, and even terminate the contract if the operator deviates from its offered processes.

Compared with scoring rules covering a few easily observable qualities, or a lowest price mechanism with minimum requirements, Västtrafik's scoring rules require substantially more work hours for bid evaluation and monitoring. As each evaluated answer is graded in two dimensions, all members of the evaluation team must decide on 26–44 individual grades for each evaluated bid, on questions ranging from vehicle supply to personnel management. Then, the evaluation team must negotiate a final grading and score for each bid. This process, which is not required when Västtrafik procures based on the lowest price, is carried out at least once more during the contract. The bidders are also required to undertake more work when their offered processes are part of the bid evaluation. For all but the winning bidder, the work of preparing a bid is a sunk cost. However, we have found that this is not reflected in the ratio between bid levels and number of vehicle kilometers. We do not have data on Västtrafik's internal procurement costs or the total final payment to the winning bidder.

Transparency and fair competition are cornerstones of public procurement. Arguably, there is a greater risk of ambiguity when evaluating subjective qualities, such as offered management processes, rather than more easily defined qualities. In 2018, the Swedish administrative court of Jönköping ruled (case no. 6154-17 judgement 2018-05-25) that the quality criteria used in a similar variant of scoring rules did not satisfy the clarity requirements of the relevant procurement law. The contract was for public railway traffic. Bidders described their offered processes by answering questions categorized under four thematic categories: *manage and lead*,

prepare delivery, support delivery, and perform delivery. The appeal was that important words in the questions and the description of the grading criteria were not clearly defined. In the judgement, the court stated that a rhetorically skilful presentation was given an undue advantage relative to qualities that were of actual significance for the procured service. Furthermore, it was stated that there was considerable risk of suppliers making different interpretations of what the procurer demanded. The court ruled that the procurement was to be redone.

A further observation is that Swedish public bus service contracts typically include several penalties or bonuses for aspects such as cancelled and delayed departures, driving style, driver behavior, and vehicle cleanliness (Vigren, 2018b). Like Västtrafik's implementations of scoring rules, these are also measures that the PTAs use to attain better quality. However, as presented in the introduction, previous studies of such incentives in Swedish tendering of public transport indicate poor results. Overall, empirical findings from Sweden suggest that it is challenging to incentivize quality and efficiency improvements in procurement, despite the level of effort and experience of the procurer.

6. Conclusions

The aim of this paper is to contribute empirical findings regarding the effectiveness of the scoring rule award mechanism used in public procurement. There are few previous studies of this matter, even though scoring rules are widely used in public procurement and, moreover, are promoted in the EU. One Swedish regional PTA has alternated between lowest price and scoring rule award mechanisms in tendering their public bus service contracts. We used this alternation in testing the effectiveness of the two award mechanisms. As this PTA's scoring rules involve the evaluation of offered processes and are intended to improve multiple areas of bus operator performance, we therefore examined the outcome in terms of performance rather than value or cost-effectiveness. Our research question for this study was: How did the use of scoring rules influence the punctuality level? Punctuality corresponds to processes evaluated in these scoring rules, and has previously been identified as an important quality in public transport.

The quantitative analysis provides no support for punctuality being influenced by the examined variants of scoring rules. This is despite the considerable effort put into formulating complex scoring rules, evaluating offers, and incentivizing performance throughout the contracts. Conversely, the results imply that the PTA has been able to attain equally good punctuality using the simpler lowest price approach. Our overall conclusion is that current promotion of the scoring rule award mechanism mainly reflects its conceptual benefits. Knowledge and practical guidance about how scoring rules should be implemented to attain better quality and value are still limited. Moreover, we observe that little guidance is available on how well scoring rule versus lowest price mechanisms are suited for different circumstances. Given the widespread use of variants of scoring rules, there should be many cases that can be used in examining these issues.

7. Acknowledgements

The authors would like to thank Andreas Vigren for sharing his tendering dataset and related documents—these form the basis of this paper, as do the data generously provided to us by helpful personnel at Västtrafik. We are very grateful for these materials. We are also grateful for the feedback and useful discussions at the Thredbo 16 conference, the review seminar on the paper at VTI, and Transportforum 2020. Our colleague Samuel Lindgren has been a great sounding board concerning the details of the empirical analysis. Finally, we thank Johan Nyström, Kristofer Odolinski, and Jan-Eric Nilsson for their valuable input throughout this research project.

8. Declaration of interest

This research project was funded by K2, Sweden's national center for research and education on public transport. K2 is operated and funded by Lund University, Malmö University, and VTI in cooperation with Region Stockholm, Region Västra Götaland, and Region Skåne. Ridderstedt's participation in the Thredbo 16 conference was partly financed by the Tower Transit Award by the operator Tower Transit.

References

- Abkowitz, M. D., & Engelstein, I. (1983). Factors affecting running time on transit routes. *Transportation Research Part A: General*, 17(2), 107-113.
- Arellano, M. (1993). On the testing of correlated effects with panel data. *Journal of econometrics*, 59(1-2), 87-97.
- Asker, J., & Cantillon, E. (2008). Properties of scoring auctions. *The RAND Journal of Economics*, 39(1), 69-85.
- Asker, J., & Cantillon, E. (2010). Procurement when price and quality matter. *The Rand journal of economics*, 41(1), 1-34.
- Australian Department of Defense (2020). Defence procurement policy manual 11 December 2020.
- Baltagi, B. H., Bresson, G., & Pirotte, A. (2003). Fixed effects, random effects or Hausman–Taylor?: A pretest estimator. *Economics letters*, 79(3), 361-369.
- Batarce, M., & Ávila, F. (2020). Misguided quality incentives: The case of the Santiago bus system. *Transportation Research Part A: Policy and Practice*, 139, 376-399.
- Bergman, M. A., & Lundberg, S. (2013). Tender evaluation and supplier selection methods in public procurement. *Journal of Purchasing and Supply Management*, 19(2), 73-83.
- Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The review of economic studies*, 47(1), 239-253.
- Börjesson, M., & Rubensson, I. (2019). Satisfaction with crowding and other attributes in public transport. *Transport Policy*, 79, 213-222.
- California Department of Transportation (2002). Guidelines for use of A+B bidding provisions.
- Che, Y. K. (1993). Design competition through multidimensional auctions. *The RAND Journal of Economics*, 668-680.
- Diab, E. I., & El-Geneidy, A. M. (2013). Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport*, 4(3), 209-231.
- Directive 2014/24/EU of the European Parliament and of the Council of 26 February 2014 on public procurement and repealing Directive 2004/18/EC. Official Journal of the European Union, 28/3/2014. L 94/65.
- Decarolis, F., Spagnolo, G., & Pacini, R. (2016). *Past performance and procurement outcomes* (No. w22814). National Bureau of Economic Research.
- European Commission (2016). *Buying green! A handbook on green public procurement*, 3rd Edn.
- Guo, Z. (2013). Does residential parking supply affect household car ownership? The case of New York City. *Journal of Transport Geography*, 26, 18-28.
- Hart, O., & Moore, J. (1999). Foundations of incomplete contracts. *The Review of Economic Studies*, 66(1), 115-138.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–1272.
- Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric Society*, 1377-1398.
- Hensher, D. A., Stopher, P., & Bullock, P. (2003). Service quality—developing a service quality index in the provision of commercial bus contracts. *Transportation Research Part A: Policy and Practice*, 37(6), 499-517.
- Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74-91.
- Holmström, B., & Tirole, J. (1989). The theory of the firm. *Handbook of industrial organization*, 1, 61-133.
- Jansson, K., & Pyddoke, R. (2010). Quality incentives and quality outcomes in procured public transport—Case study Stockholm. *Research in Transportation Economics*, 29(1), 11-18.
- Kain, J. F., & Liu, Z. (1996). An econometric analysis of determinants of transit ridership: 1960-1990. United States. Federal Highway Administration.
- Kitamura, R. (1989). A causal analysis of car ownership and transit use. *Transportation*, 16(2), 155-173

- Koning, P., & Van De Meerendonk, A. (2014). The impact of scoring weights on price and quality outcomes: An application to the procurement of Welfare-to-Work contracts. *European Economic Review*, 71, 1-14.
- Lewis, G., & Bajari, P. (2011). Procurement contracting with time incentives: Theory and evidence. *The Quarterly Journal of Economics*, 126(3), 1173-1211.
- Liu, Z. (1993). Determinants of Public Transit Ridership: Analysis of Post World War II Trends and Evaluation of Alternative Networks. Cambridge, MA: Harvard University.
- Mazloumi, E., Currie, G., & Rose, G. (2010). Using GPS data to gain insight into public transport travel time variability. *Journal of Transportation Engineering*, 136(7), 623-631.
- Minnesota Department of Transportation (2005). Innovative contracting guidelines.
- Morrall, J, and Bløger, D. (1996). The Relationship Between Downtown Parking Supply and Transit Use. *ITE Journal*, 66(2), 32-36.
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69-85.
- New York State Department of Transportation (1999). Guidelines for the use of time-related contract provision.
- OECD (2017). Government at a Glance 2017, OECD Publishing, Paris
- Pyddoke, R. (2020). Penalties as incentives for punctuality and regularity in tendered Swedish public transport. *Research in transportation economics*, 83, 100948.
- Ridderstedt (2021). Implementations of the scoring rule award mechanism in Swedish tendering of public bus service [unpublished manuscript]. Transport Economics. The Swedish National Road and Transport Research Institute.
- Schaffer, M. E., Stillman, S. (2010). xtoverid: Stata module to calculate of overidentifying restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor. Retrieved from <http://ideas.repec.org/c/boc/bocode/s456779.html>
- Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal*, 13(1), 65-76.
- Shoup, D. (2005). The high cost of free parking. Chicago: Planners Press – American Planning Association.
- Statistics Sweden (2020a). Folkmängden den 1 november efter region, kön och år. Retrieved from http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START_BE_BE0101_BE0101A/FolkmandNov/table/tableViewLayout1/#
- Statistics Sweden (2020b). Folkmängd, topp 50. 31 december 2019. Retrieved from <https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/pong/tabell-och-diagram/topplistor-kommuner/folkmandg-topp-50/>
- Serman, B. P., & Schofer, J. L. (1976). Factors affecting reliability of urban bus services. *Journal of Transportation Engineering*, 102, 147-159.
- Strathman, J. G., Dueker, K. J., Kimpel, T., Gerhart, R., Turner, K., Taylor, P., ... & Hopper, J. (1999). Automated bus dispatching, operations control, and service reliability: Baseline analysis. *Transportation Research Record*, 1666(1), 28-36.
- Stilger, P. S., Siderius, J., & Raaij, E. M. V. (2017). A comparative study of formulas for choosing the economically most advantageous tender. *Journal of public procurement*, 17(1), 89-125.
- Susilo, Y. O., & Cats, O. (2014). Exploring key determinants of travel satisfaction for multi-modal trips by different traveler groups. *Transportation Research Part A: Policy and Practice*, 67, 366-380.
- Swedish Institute for Quality (2017). SIQs Modell för Kundorienterad Verksamhetsutveckling.
- Swedish Meteorological and Hydrological Institute (2019a). Meteorologiska observationer - Lufttemperatur, medelvärde 1 dygn [data set]. Retrieved from <https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=all,stationid=71420>
- Swedish Meteorological and Hydrological Institute (2019b). Meteorologiska observationer - Nederbörd, summa 1 dygn [data set]. Retrieved from <https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=precipitation24HourSum,stations=all,stationid=71420>
- Taylor, B. D., & Fink, C. N. Y. (2013). Explaining transit ridership: What has the evidence shown?. *Transportation Letters*, 5(1), 15-26.
- Tender Electronic Daily (2019). TED – Contract award notices 2010-2019. Retrieved from <https://data.europa.eu/euodp/en/data/dataset/ted-csv>
- Texas Department of Transportation (2018). Accelerated construction guidelines.
- U.S. Government Accountability Office (2017). Defense contracting. DOD's use of lowest price technically acceptable source selection procedures to acquire selected services. GAO-18-139.

- van Lierop, D., Badami, M. G., & El-Geneidy, A. M. (2018). What influences satisfaction and loyalty in public transport? A review of the literature. *Transport Reviews*, 38(1), 52-72.
- Vigren, A. (2018a). Tendering data [data set].
- Vigren, A. (2018b). How many want to drive the bus? Analyzing the number of bids for public transport bus contracts. *Transport Policy*, 72, 138-147.
- Vigren, A. (2020). The Distance Factor in Swedish Bus Contracts: How far are operators willing to go? *Transportation Research Part A: Policy and Practice*, 132, 188-204.
- Vigren, A., & Pyddoke, R. (2020). The impact on bus ridership of passenger incentive contracts in public transport. *Transportation Research Part A: Policy and Practice*, 135, 144-159.
- Västtrafik (2011). Trafiktjänster Buss 2013: Tvåstad [tendering documents].
- Västtrafik (2013). Trafik 2014 Skövde, Skaraborg Express, Uddevalla och Borås [tendering documents].
- Västtrafik (2014). Rekommendation för beslut om allmän trafikplikt Upphandling Trafik 2016. 2014-10-22. Retrieved from <https://www.svenskkollektivtrafik.se/globalassets/partnersamverkan/dokument/om-oss/dokumentation-konferenser/kollas-arskonferens-2015/rekommendation-for-beslut-om-allman-trafikplikt-trafik-2016.pdf>
- Västtrafik (2015). Trafik 2016 Express Fyrbodalen, Stenungsund och Tjörn Express, Tjörn, Röd och Lila Express [tendering documents].
- Västtrafik (2020a). Årsredovisning 2019. Västtrafik Ab [Annual report].
- Västtrafik (2020b). E-mail correspondence with procurer at Västtrafik. E-mails received 2020-03-10 and 2020-03-16.
- Waara, F., & Bröchner, J. (2006). Price and nonprice criteria for contractor selection. *Journal of construction engineering and management*, 132(8), 797-804.
- Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data MIT press. *Cambridge, MA*, 108.

Appendix

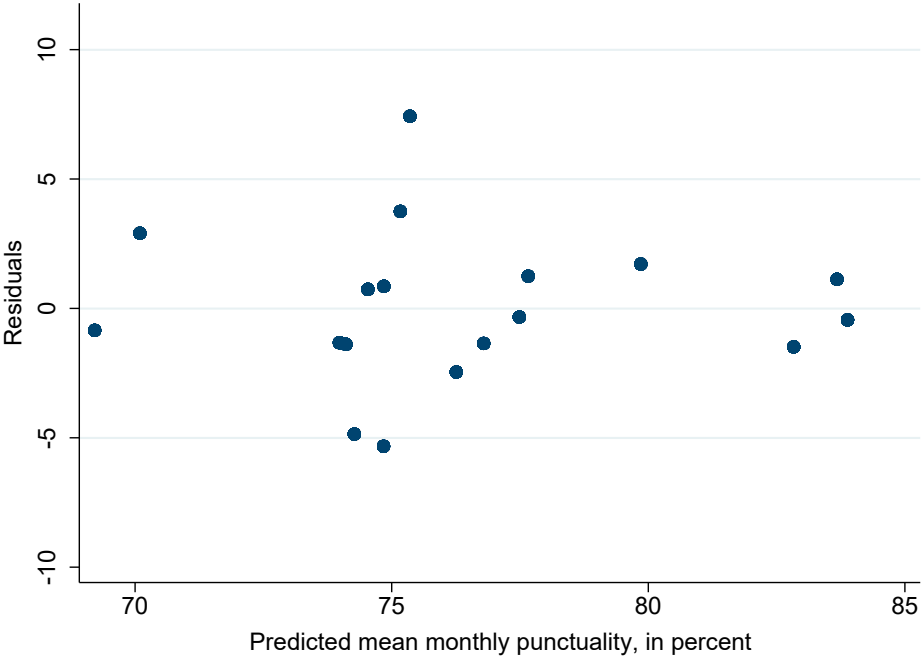


Fig. A1. Distribution of the residuals: OLS (Table 7, column 2)

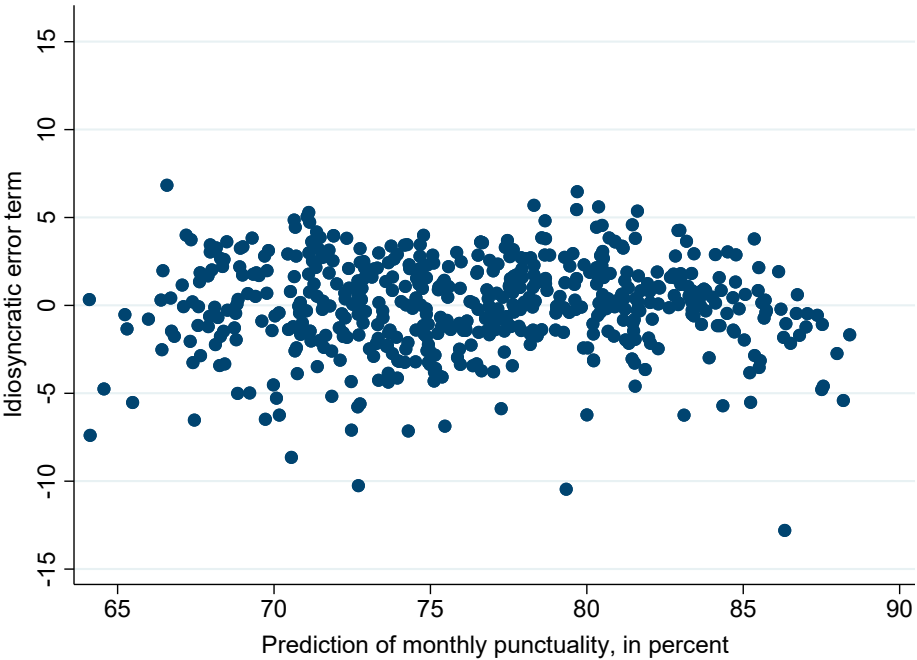


Fig. A2. Distribution of the residuals: CRE (Table 7, column 7).

Table A1 Results table. Sensitivity analysis with interactions.

	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
	CRE	CRE	CRE	CRE	CRE	CRE	CRE	CRE
SRULE	2.99 (0.90)	-4.37 (-0.60)	-0.60 (-0.18)	0.23 (0.08)	4.59 (1.23)	-0.60 (-0.18)	-11.17** (-2.18)	1.46 (0.60)
In(PASS)	-4.87** (-2.56)	-6.41*** (-5.94)	-6.41*** (-5.94)	-6.41*** (-5.94)	-6.41*** (-5.94)	-6.41*** (-5.94)	-6.41*** (-5.95)	-5.98*** (-5.24)
VKM	-2.57** (-2.35)	-4.61 (-1.28)	-1.92*** (-2.66)	-1.64*** (-3.36)	-2.04*** (-3.71)	-1.92*** (-2.66)	-2.14*** (-4.36)	-1.77*** (-3.18)
CITY	7.49*** (3.35)	6.47** (2.14)	6.04 (1.49)	6.31** (2.42)	8.26*** (3.20)	7.30** (2.49)	4.70 (1.32)	6.64** (2.49)
ROUTES	0.08 (1.32)	0.11 (1.49)	0.11* (1.65)	0.21 (1.44)	0.07 (1.12)	0.11* (1.65)	0.10 (1.50)	0.12* (1.80)
STOPS	-0.19** (-2.14)	-0.20** (-2.09)	-0.15 (-1.63)	-0.14 (-1.58)	-0.09 (-0.99)	-0.15 (-1.63)	-0.21*** (-2.83)	-0.15 (-1.62)
DPEN	-0.72 (-0.26)	-0.78 (-0.31)	0.88 (0.37)	1.06 (0.43)	-0.62 (-0.25)	-0.37 (-0.07)	1.01 (0.40)	1.01 (0.40)
EXPER	-0.11* (-1.92)	-0.10*** (-2.61)	-0.08** (-1.97)	-0.07*** (-2.70)	-0.09*** (-2.88)	-0.08** (-1.97)	-0.26*** (-3.31)	-0.08*** (-3.06)
Constant	95.79*** (9.15)	101.32*** (6.50)	91.18*** (9.75)	87.99*** (14.52)	90.28*** (13.77)	91.18*** (9.75)	107.62*** (11.06)	88.38*** (14.39)
<i>Interactions with SRULE</i>								
	PASS	VKM	CITY	ROUTES	STOPS	DPEN	EXPER	TEMP
	-2.37 (-0.94)	2.32 (0.74)	1.26 (0.25)	-0.15 (-1.09)	-0.10 (-0.59)	1.26 (0.25)	0.17** (2.45)	-0.03 (-0.51)
								PRECIP -0.98* (-1.94)
								TE*PR 0.04 (1.47)
<i>Controls:</i>								
Weather	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	540	540	540	540	540	540	540	540
R ² : Betw.	0.67	0.67	0.65	0.65	0.68	0.65	0.76	0.65

Dependent variable: Monthly punctuality $\in [0, 100]$ (i.e. share of monitored departures from bus stops less than 30 seconds early and 3 minutes late). Cluster robust t statistics in parentheses. Clusters (15): Traffic assignments tendered as part of the same contract. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

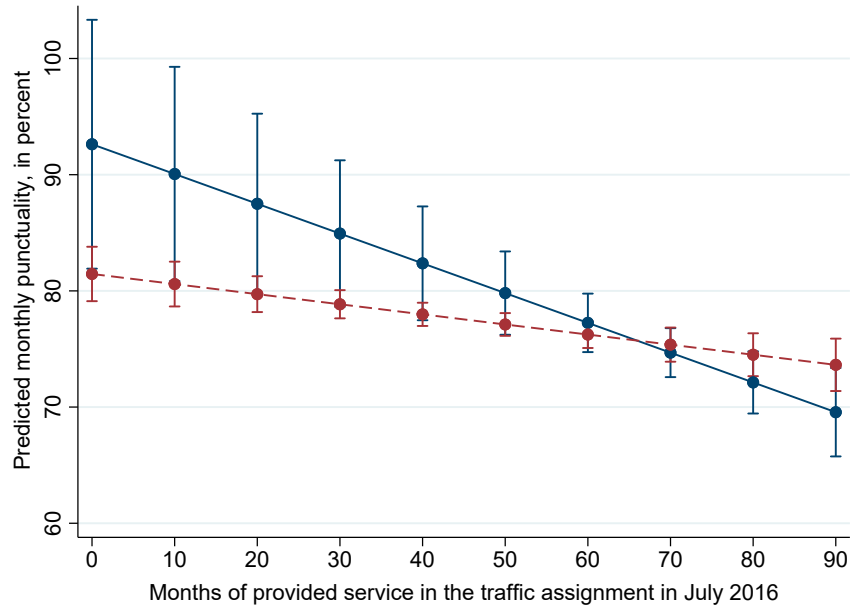


Fig. A3. The combined effect of scoring rules and months of provided service in a traffic assignment at the start of the studied period (Table A1, column 8). Blue solid line: lowest price. Red dashed line: scoring rules. Confidence intervals at the 95% level.

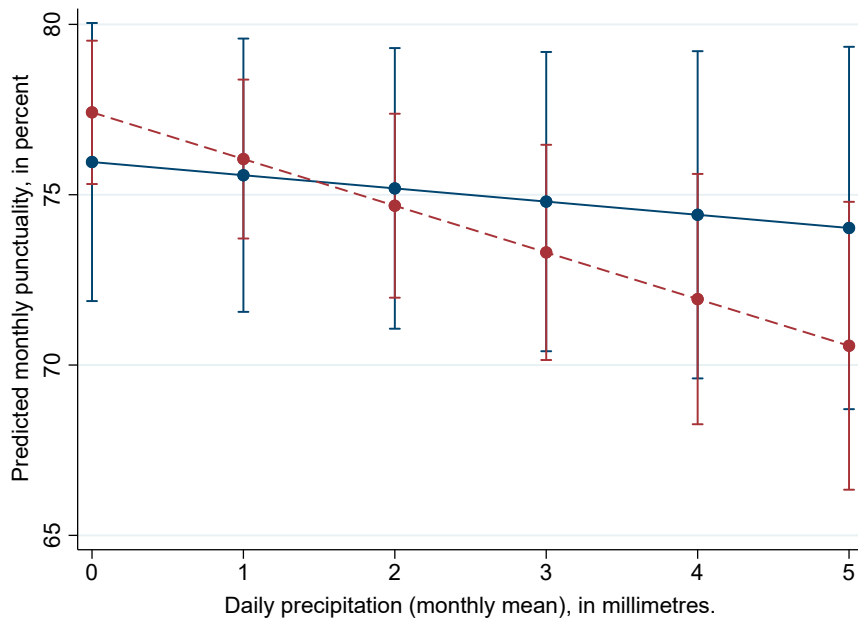


Fig. A4. The combined effect of scoring rules and monthly mean of daily precipitation (Table A1, column 9). Temperature set to zero. Blue solid line: lowest price. Red dashed line: scoring rules. Confidence intervals at the 95% level.



K2 is Sweden's national centre for research and education on public transport. This is where academia, the public sector and industry meet to discuss and develop the role of public transport.

We investigate how public transport can contribute to attractive and sustainable metropolitan areas of the future. We educate members of the public transport sector and inform decision-makers to facilitate an educated debate on public transport.

K2 is operated and funded by Lund University, Malmö University and VTI in cooperation with Region Stockholm, Region Västra Götaland and Region Skåne. We receive financial support from Vinnova, Formas and the Swedish Transport Administration.

www.k2centrum.se/en

