Journal of Public Economics 157 (2018) 78–94

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Bunching on the Autobahn? Speeding responses to a 'notched' penalty scheme☆

Christian Traxler[a], Franz G. Westermaier[b], Ansgar Wohlschlegel[c]

[a] Hertie School of Governance, Max Planck Institute for Research on Collective Goods, and CESifo, Germany
[b] University of Wuppertal, Germany
[c] University of Portsmouth, United Kingdom

ARTICLE INFO

Keywords:
Notches
Speeding tickets
Stepwise penalty function
Salience
Simplicity

JEL classification:
K4
D04
R41

ABSTRACT

This paper studies drivers' responses to a 'notched' penalty scheme in which speeding penalties are stepwise and discontinuously increasing in speed. We present survey evidence suggesting that drivers in Germany are well aware of the notched penalty structure. Based on a simple analytical framework, we analyze the impact of the notches on drivers' optimal speed choices. The model's predictions are confronted with data on more than 150,000 speeding tickets from the Autobahn and 290,000 speed measures from a traffic monitoring system. The data provide evidence on modest levels of bunching, despite several frictions working against it. We analyze the normative implications and assess the scope for welfare gains from moving from a simple, notched penalty scheme to a more complex but less salient Pigouvian scheme.

1. Introduction

The economics of externality-correcting interventions typically focuses on smooth, Pigouvian incentive schemes. It is quite common, however, for such subsidies, taxes, or penalty schemes to include 'kinks' or 'notches' (Slemrod, 2013; Kleven, 2016). In North America and some continental European countries, for instance, vehicles are subject to a tax that is a step function of a car's fuel economy or CO_2 emission level (Sallee and Slemrod, 2012; D'Haultfœuille et al., 2016). Notches are also ubiquitous in many domains of law enforcement. Fines and penalties often change discontinuously with small variations in the 'nuances' of a legal violation. Compared to minor fraud, theft, or tax evasion, major cases are punished in very different ways — and the differences between minor and major are commonly defined along cutoffs in a continuous metric of damage (Rasmusen, 1995). A further example is driving under the influence, which triggers a penalty that discontinuously increases at certain cutoff levels of blood alcohol content (Hansen, 2015). Similarly, many countries use penalties for speeding which are stepwise increasing in the speed level (Goncalves and Mello, 2017).

From a Pigouvian perspective, such notched taxes or enforcement schemes seem puzzling. Given that the underlying externalities are typically smooth functions of a given action, corrective interventions should build on smooth incentives, too. The marginal cost imposed by a Pigouvian scheme should reflect the marginal social cost of the action — and it is unclear why the latter should discontinuously change at certain cutoffs (such as, e.g., speeding at 30 mph above the limit). Admittedly, the exact marginal externality may depend on a vast number of factors.[1] A 'correct' Pigouvian solution would thus become quite complex. A high level of complexity, in turn, might hamper – in this case *desired* – behavioral responses to the policy (e.g., Abeler and Jäger, 2015). Notched Pigouvian schemes could, therefore, represent constrained approximations to the optimal penalties that account for decision makers' limitations in their attention, awareness, or cognition (Altmann et al., 2017; Ericson, 2011; Taubinsky and Rees-Jones, 2017). The simplicity of, e.g., stepwise increasing penalty schemes might assure its salience and individuals' awareness about the scheme (Chetty et al., 2009; Finkelstein, 2009). In the presence of externalities, the salience gains related to simplicity could then dominate the welfare costs from deviating from the correct Pigouvian scheme.

We examine this idea in the context of speed limit enforcement in Germany, where drivers are confronted with a notched – and, as we are

☆ We are grateful to two anonymous referees and to Camille Landais for many helpful suggestions. We would also like to thank Henrik Kleven, Håkan Selin, Joel Slemrod, and numerous conference and seminar participants for fruitful comments. Marie Baumann, Lena Föhrenbacher, Niclolai Vitt and Tarek Zayed provided excellent research assistance. Traxler acknowledges financial support from the German Research Foundation (DFG grant TR1471/1).

E-mail addresses: traxler@hertie-school.org (C. Traxler), westermaier@uni-wuppertal.de (F.G. Westermaier), Ansgar.Wohlschlegel@port.ac.uk (A. Wohlschlegel).

[1] Consider, e.g., the marginal accident risks associated with a marginal increase in speeding under different weather, road, traffic and car conditions.

https://doi.org/10.1016/j.jpubeco.2017.11.006
Received 20 January 2016; Received in revised form 21 November 2017; Accepted 27 November 2017
Available online 05 December 2017

going to show, quite salient – penalty scheme. The core of our analysis exploits detailed data on more than 150,000 speeding tickets recorded on the German Autobahn. In addition, we examine complementary data on 290,000 (non-ticketed) speed measures from single-lane roads. Like in many other countries, fines and other penalties jump discontinuously at several speed levels. Hence, drivers face numerous 'notches' in the penalty structure.

To set the stage for our analysis, we first introduce a simple analytical framework in the spirit of Kleven and Waseem (2013) and study the effect of the penalty notches on drivers' optimal speed choices. The analysis, which differs in several ways from a deterministic taxation framework, highlights conditions under which drivers bunch at different notches. Before examining bunching responses, we present evidence from a survey that assesses whether drivers know and understand the penalty scheme's structure. The survey reveals that respondents are quite knowledgeable about the scheme's stepwise shape, its discontinuous jumps and the location of the cutoffs. This finding is by no means trivial and – potentially due to the penalty scheme's simplicity – different from studies documenting limited knowledge (e.g., Chetty et al., 2013) and misconceptions of non-linear or non-convex budget sets (De Bartolome, 1995, Liebman and Zeckhauser, 2004; Feldman et al., 2016).[2] At the same time, the survey indicates considerable heterogeneity across drivers, suggesting that bunching might not be very sharp. Variation in the speed indicated by speedometers as well as ambiguities regarding the computation of the fine-relevant speed measure might further limit the scope for bunching.

Studying the distribution of measured speed, we do find bunching for some of the notches in the penalty scheme. Disproportionately more drivers are speeding exactly at (or slightly below) certain cutoffs of the penalty scheme. Quantitatively, however, the observed bunching mass is quite small and – similar to findings in several tax bunching studies (Saez, 2010; Chetty et al., 2011; Bastani and Selin, 2014) – varies considerably along the speed distribution.[3] Given the optimization frictions noted above, it is not too surprising that we do not detect massive bunching responses. The fact that we do find some evidence on bunching indicates that at least some drivers do respond despite the frictions they face. By exploring a reform of the penalty system we provide further evidence on behavioral responses. The reform increased the size of several notches, without shifting their location. Our results indicate that some speeders responded to the reform by avoiding speed ranges which triggered significantly higher penalties after the reform. Overall, the data suggest that the reform produced a sizable shift in the speeding distribution, with a 25% drop in the fraction of vehicles driving more than 20 km /h above the limit.

Based on our empirical results we finally get back to the analytical framework and derive an upper bound on the 'salience gap' — the maximum difference in salience between a notched and a more complex Pigouvian penalty scheme, for which the latter scheme would be superior to the notched one. Our results suggest that if at most 12% of drivers turn inattentive after switching from a notched to a smooth Pigouvian scheme, the hypothetical reform would be welfare enhancing. For a higher salience gap, the notched approximation would be superior to a Pigouvian scheme that is poorly understood by drivers.

Our study relates to several strands of research. First, we contribute to the law and economics of speeding and speed control.[4] One million lives are lost worldwide each year due to motor vehicle accidents (Peden et al., 2004), with speeding being a major contributor to the number of traffic fatalities. It is therefore important to advance our understanding of speed control policies. Several quasi-experimental studies document the impact of police enforcement (DeAngelo and Hansen, 2014), speed limits (van Bentham, 2015) and speeding tickets (Dusek and Traxler, 2017) on travel speed, accidents, fatalities and air pollution externalities. The present study differs from these contributions since it analyzes drivers' responses to the specific *structure* of speeding penalties. By focusing on an electronic (i.e., automated) enforcement system, our study also differs from Goncalves and Mello (2017), who study racial bias in police officers' (not drivers') responses to similar notches.

Conceptually, our paper closely relates to the work by Sallee and Slemrod (2012), who study automakers' responses to notches in the (fuel economy based) taxation of automobiles. They offer an interesting welfare discussion, in which they quantify the welfare loss due to the inaccurate incentives of a notched scheme. Our analysis contributes to this discussion by (i) explicitly modeling the benefit of a notched scheme being more salient and by (ii) empirically approximating the tradeoff between these costs and benefits. In this vein, we also offer a new perspective on related welfare discussions of notched schemes (e.g., Blinder and Rosen, 1985; Gillitzer et al., 2017).

In terms of methods, we use tools from public finance to analyze behavioral responses to law enforcement (Kleven, 2016). Applying the bunching framework to speeding responses, our analysis clarifies several key differences between the law enforcement and the taxation context. In our application, bunching is proportional to *expected* notches — the discontinuous increase in penalties weighted with the detection probability. Hence, there are two policy parameters that jointly determine the incentive to choose a corner solution: the jump in penalties at a given speed (analogous to, e.g., increases in average tax rates at certain income levels) and the risk of punishment. This latter dimension, which is not present in most taxation studies but crucial if one explores notches in law enforcement, impedes the translation of bunching mass into behavioral response elasticities. The reason is that objective variation in law enforcement and subjective priors about detection risks (speed controls) essentially add an additional layer of heterogeneity. Without common knowledge about the enforcement parameters (or subjective risk assessment data), notches in penalty schemes cannot readily be used to identify behavioral elasticities.[5]

The rest of the paper is structured as follows. Section 2 describes the institutional framework for speeding in Germany. Section 3 presents evidence from our survey on drivers' knowledge of the penalty scheme's structure. Section 4 introduces a simple model of speeding and discusses several predictions. After presenting the data (Section 5), we turn to the empirical analysis of the different speed measurement data (Section 6). Section 7 offers a welfare comparison between a notched and a smooth penalty scheme and Section 8 concludes.

2. Institutional background

Despite common prejudices about German highways being the great dream of speeders, there are speed limits on more than 85% of the 13,000 km of Autobahn. Speed limits are primarily imposed for safety reasons: high speed is the leading cause of roughly 4000 annual traffic deaths and 400,000 annual traffic injuries in Germany. For most of the accidents, speeding is the chief cause.[6]

The enforcement of speed limits is based on permanently installed and on mobile speed cameras, which are set up by an officer for a few

[2] Further evidence is discussed in Congdon et al. (2011; Ch. 2).

[3] Using the data that cover the full distribution of measured speed (rather than the distribution conditional on receiving a speeding ticket), for instance, we find no evidence on bunching at the speed limit itself. Moreover, in both sets of data, we observe no bunching at very high speed level cutoffs. Drivers in this speed range might either have a very 'sharp' preference for speeding or they might expect very low detection probabilities. In either case, such drivers seem fairly insensitive to the penalty notches.

[4] For early, theoretical contributions in this field see, e.g., Jondrow et al. (1983), Lave (1985), and Graves et al. (1993).

[5] A further, more technical difference to the taxation literature is related to the close proximity of potential bunching points. In our context, it is reasonable to consider drivers who are indifferent about speeding 20 or 25 km/h above the limit. Our analysis therefore considers the joint influence of multiple, potentially inter-related notches on behavior — a point which advances and generalizes the theoretical bunching literature.

[6] See *Deutsches Statistisches Bundesamt* (2017; *Fachserie 8/7, Verkehr*).

Table 1
Penalties for speeding (outside built-up areas).

Speed bracket	Cutoff	Fines (in euro)					Penalty points	Driving ban
(speed above limit in km/h)		Pre-reform		Post-reform		Change		(Months)
	x_i	$\tilde{f}_i^{\text{pre}}$	$\tilde{\Delta}_i^{\text{pre}}$	$\tilde{f}_i^{\text{post}}$	$\tilde{\Delta}_i^{\text{post}}$	$\frac{\tilde{\Delta}^{\text{post}}-\tilde{\Delta}^{\text{pre}}}{\tilde{\Delta}^{\text{pre}}}$		
$x \leq 10$	10	10.0	10.0	10.0	10.0	0	0	–
$10 < x \leq 15$	15	20.0	10.0	20.0	10.0	0	0	–
$15 < x \leq 20$	20	30.0	33.5	30.0	63.5	0.90	0	–
$20 < x \leq 25$	25	63.5	10.0	93.5	10.0	0	1	–
$25 < x \leq 30$	30	73.5	15.0	103.5	40.0	1.67	3	(1)
$30 < x \leq 40$	40	98.5	25.0	143.5	40.0	0.60	3	(1)
$40 < x \leq 50$	50	123.5	50.0	183.5	80.0	0.60	3	1
$50 < x \leq 60$	60	173.5	125.0	263.5	200.0	0.60	4	1
$60 < x \leq 70$	70	298.5	100.0	463.5	160.0	0.60	4	2
$70 < x$	–	398.5	–	623.5	–	–	4	3

Notes: The table presents the fines ($\tilde{f}_i$) and other penalties (penalty points and temporary driving bans) for different speed levels over the speed limit on the German Autobahn. The columns labeled by $\tilde{\Delta}^{\text{pre}}$ and $\tilde{\Delta}^{\text{post}}$ capture the pre-/post-reform notches in the fines, respectively (i.e., the increase in the fine associated with moving from a 'lower' to a 'higher' speed bracket). The notation uses tilde to indicate that these variables (i.e., $\tilde{f}_i$ and $\tilde{\Delta}_i$) only refer to the monetary component of the penalty. In the model section, f_i and Δ_i refer to the disutility from all different types of penalties for a given speed x. The duration of temporary driving bans indicated in brackets are only imposed the second time a driver is detected speeding by more than 25 km/h within one year.

hours. (On-board speed measurement in unmarked cars are extremely rare.) If a vehicle passes with a speed above a certain enforcement level (see Section 5 below), a picture is automatically taken and the speed is recorded. The vehicles' owners are identified from the number plates and receive a ticket by mail. Penalties for speeding offenses – i.e., fines, 'penalty points' and possible driving bans – are a function of the measured speed: the speed camera's measure s is first rounded down to the next integer; a tolerance value of 3% is subtracted and the result is again rounded to the next lower integer.[7] The outcome from this so-called 'tolerance rule' (which serves as a concession to prevent appeals against speeding tickets), the speed level x, determines the penalty according to Table 1.

Monetary fines range from 10 to 623.50 euro. The fines discontinuously increase at cutoffs of x being, e.g., 20, 25, 30 or 40 km/h above the speed limit. Within each speed bracket, i.e., between two cutoffs, the fine is constant. The same holds for temporary driving bans: a one-, two- or three-month ban is imposed for speeding in the range 40–60, 60–70, and more than 70 km/h above the limit, respectively. The penalty point scheme follows a stepwise pattern, too. Speeding in the range 20–25, 25–50, or above 50 km/h results in one, two or three penalty points, respectively. The repeated accumulation of points, which are recorded in a register of traffic offenders, can result in the revocation of a driver's license.[8]

In 2009, there was a significant reform of the penalty schedule. Starting with February 2009, fines for speeding with more than 20 km/h above the limit were increased considerably. All other penalties (points and driving bans) remained unchanged. An overview of the fines before and after the reform as well as the stable components of the penalty system is provided in Table 1.

The table illustrates the key property of the penalty system: the penalty scheme is characterized by what the Public Finance literature calls 'notches' (Slemrod, 2013): discontinuous increases in fines, penalty points and/or driving bans at each speed bracket's cutoff.[9] Before 2009, speeding with e.g. 20 km /h above the limit triggered a fine of 30 euro and no penalty point; for 21 km /h, it was 63.50 euro and one penalty point. The table further shows that the reform not only increased the level of the fines but also the magnitude of several notches. At the 20 km /h cutoff, for instance, the increase in the fine amounted to 33.50 euro before ($\tilde{\Delta}^{\text{pre}}$) but 63.50 euro after the reform ($\tilde{\Delta}^{\text{post}}$). The notch in fines thus increased by 90%. Similar increases occurred at other cutoffs. Further dimensions of penalties beyond fines (i.e., driving bans, penalty points) remained the same.

3. Survey evidence

Let us first study whether the simplicity of the stepwise penalty structure is reflected in a good knowledge of the penalty scheme. To approach this question, we conducted an online survey. The survey was implemented in June 2013 in cooperation with a professional survey company which maintains a large sample of German individuals that is representative in several dimensions (age, gender, education and occupational structure). We invited a random subset of this sample (conditional on having a driver's license) to participate in the survey. The key survey questions and summary statistics for the approximately 1000 participants are provided in the Online Appendix (see Table A.1). 48% of the respondents are male and the average age is 43 years. 54% drive on daily basis and more than 60% drive on the Autobahn several times a month. Over a third of the respondents have experience with the penalty system: 28% report that they were caught speeding during the last two years and 12% indicate that they hold a positive penalty point record in the register of traffic offenders (see above).

To elicit people's knowledge about the penalty system, we first asked survey participants to indicate the level of speeding fines for a *randomly drawn* sequence of speed levels (see Online Appendix). Comparing answers within and between respondents, we thus obtain information on the expected fines as a function of the speed *without* mentioning the stepwise fine structure in the question. It is important to note that this way of asking the question (i.e., by exposing subjects to randomly drawn speed levels) gives the survey a hard time to reveal a

[7] Consider the following example: a speed camera measures $s = 140.6$ km/h. The recorded speed is first rounded down to 140 km/h. Thereafter, it is reduced by 3% to 135.8 km/h and further rounded down to $x = 135$ km/h. Note that a similar (but informal and less institutional) rounding in favor of the speeder is frequently applied in institutional contexts that include face to face interactions with police officers.

[8] Offenders who have accumulated between 14 and 17 points are obliged to participate in a costly seminar on traffic safety. Drivers that end up at 18 or more points get their driver's license revoked. Older points are deleted two years after collecting them if no additional tickets were issued since then. For a theoretical analysis of combining monetary fines with penalty points, see Bourgeon and Picard (2007).

[9] Notched penalty structures can be found in many other countries. For evidence from Italy, Spain, and the Czech Republic, see De Paola et al. (2013), Castillo-Manzano et al. (2010) and Montag (2014), respectively.

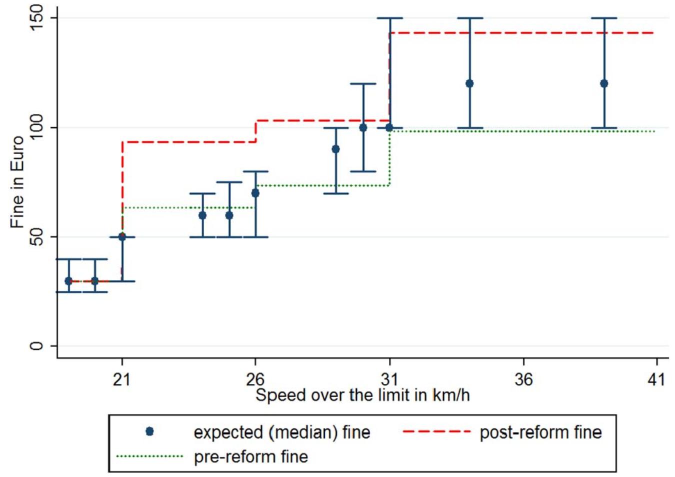


Fig. 1. Expected and actual fines (in euro) for a given speed above the limit. *Note:* The figure illustrates survey responses regarding the expected fines (in euro, vertical axis) for a given speed above the limit (in km/h, horizontal axis). The blue dots capture median expectation, the upper and lower 'bounds' on the blue dots indicate the 33rd- and 66th-percentile, respectively. The dashed red lines and the green dotted line depict the fines for the post- and pre-reform period, respectively. (See the Online Appendix, Survey Question 1.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

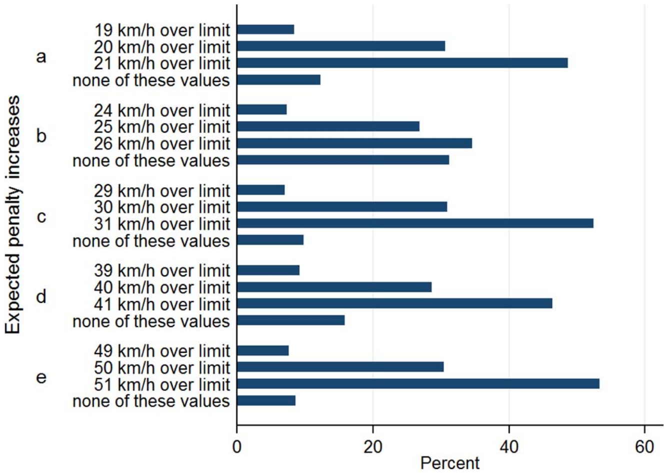


Fig. 2. Expectations regarding the cutoff points. *Notes:* For each of the five cutoffs (21, 26, 31, 41 and 51 km/h above the speed limit), the bar graphs represent the fraction of respondents who expect an increase in the penalties at one (or none) of the indicated speed levels. (See the Online Appendix, Survey Question 2.)

clear step function in the expected fines. Moreover, the randomization approach implies that Fig. 1 – which illustrates the results – is not simply comparing answers for a given set of speed levels but it pools variation within and between respondents.

In the light of these qualifications, Fig. 1 offers a relatively clear picture. The blue dots indicate the median expected fines (together with the 33rd- and 66th-percentile) for the surveyed levels of speeding. The dashed red (dotted green) line shows the actual fine for the post- (pre-) reform period. The figure provides two insights. First, respondents reveal a good knowledge of the stepwise structure and the increase of fines at the cutoffs.[10] While there is no jump in the median response at the 30/31 threshold, the lower and top terciles (as well as the average, not depicted) strongly increase at the cutoff. Second, the expected *level* of the speeding fines is much closer to the pre-reform level. A possible interpretation of this observation is that drivers' expectations converge only slowly to the post-reform levels.

A relatively good understanding of the stepwise shape of the penalty function is further captured in Fig. 2. The graph displays responses to a more direct question that explicitly asked whether there is an increase in penalties at a certain speed level. For each of the five surveyed thresholds, the mode of the response distribution – typically accounting for half of all answers – coincides with the true cutoff. This corroborates the first finding from above.

We also asked whether drivers know about the tolerance rule for computing the speed level which is relevant for determining the penalty (compare Section 2). 93% answered that they were aware of this rule. Among them, 36% – again the mode of the response distribution – indicated the correct rule (see Online Appendix). Hence, there is quite some variation in the expectations regarding the tolerance rule, but one out of three drivers seems to perfectly know the rule.

Overall, the survey suggests that the simplicity of the penalty scheme is reflected in a relatively good understanding of the system. The majority of respondents understand well the scheme's stepwise shape with its discontinuous changes at cutoffs. This finding is by no means trivial and – potentially due to the simplicity of the penalty structure – sets it apart from a growing body of evidence on individuals' limited knowledge and misperception of non-linear budget sets (e.g., De Bartolome, 1995; Liebman and Zeckhauser, 2004; Chetty et al., 2013; Feldman et al., 2016).

The solid understanding of the notches observed in our set-up implies that drivers should respond to the penalty structure. In fact, we asked participants if they would speed on highways and, if at all, whether they would try to avoid higher fines by staying under a certain cutoff. Among those that admitted speeding (almost three out of four respondents), 85% indicated they would drive at a speed level slightly below one of the cutoffs in the penalty scheme. It is important to stress, however, that respondents' knowledge is far from being perfect. In fact, due to misconceptions regarding the location of the cutoff points and the generosity of the tolerance rule, drivers might simply target the 'wrong' speed levels.

4. Theoretical framework

To set the stage for our empirical analysis, we analyze a risk neutral driver's optimal speeding response to a stepwise penalty scheme. Let the monetary equivalent of the net benefits from a given speed x (time spent on the trip, net of costs for fuel consumption, experienced 'pleasure' from driving at speed x, etc.) be given by a twice differentiable function $v(x,\theta)$, where the parameter θ captures heterogeneous preferences. Drivers' types θ are distributed continuously with density $g(\theta)$ and the c.d.f. $G(\theta)$. For every type θ, $v(.,\theta)$ is concave in x and satisfies the single-crossing property, i.e. $\frac{\partial^2 v(x,\theta)}{\partial x^2} < 0$ and $\frac{\partial^2 v(x,\theta)}{\partial x \partial \theta} > 0 \;\forall\, x, \theta.$[11] With probability p, the driver's speed is measured by a speed camera. In this case, he may get a penalty $f(x)$ which is a step function of the observed speed x:

$$f(x) = \begin{cases} f_0 = 0, & \text{if } x \leq x_0, \text{ with } x_0 \text{ capturing the speed limit;} \\ \dots & \\ f_i, & \text{if } x_{i-1} < x \leq x_i, \\ \dots & \\ f_I, & \text{if } x_{I-1} < x, \end{cases} \tag{1}$$

with x_i denoting the cutoff for speed bracket $i = 1,\dots,I$, and f_i expressing the costs of the penalties for a given speeding bracket i.[12] A notch at a

[10] The average survey duration, as well as the question specific response times indicate that responses do *not* stem from ad-hoc online research on the questions.

[11] For a less stylized model of speeding choices (in which, however, drivers' preferences are homogeneous) see Jondrow et al. (1983).

[12] As the penalty may include non-monetary components (e.g., a driving ban or penalty points), $f(x)$ denotes the average present value of the monetary equivalent of the penalty. Allowing for heterogeneity in the penalty across different drivers would complicate the following discussion without yielding additional insights.

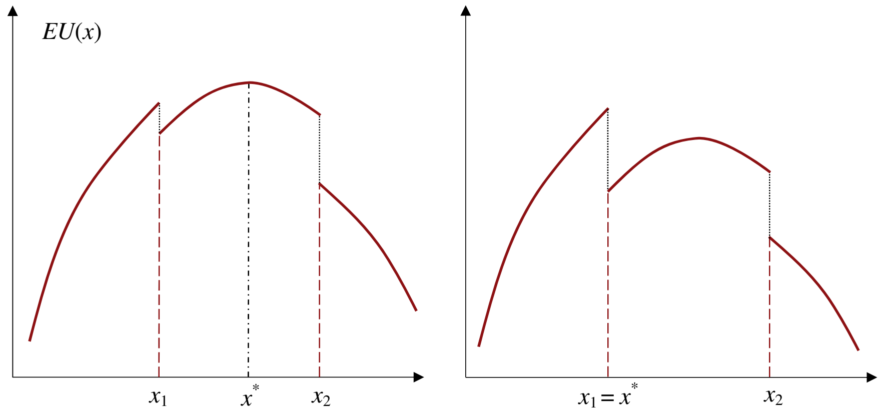


Fig. 3. Optimal speed level with notches: Interior optimum and corner solution. *Note:* The figure displays the mapping of speed x into expected utility, *EU*, for a given θ-type and notches at two cutoffs, x_1 and x_2. For the case depicted in the left panel, the expected notch is small and the driver's optimal speed corresponds to the interior solution. In the right panel, the expected notch at x_1 is larger, thus turning this cutoff into the driver's optimal speed level.

cutoff speed x_i is given by $\Delta_i := f_{i+1} - f_i > 0$. Assuming that the drivers' utility functions are quasi-linear, their objective functions are given by the net benefits from driving $v(x,\theta)$ and the expected penalties $f(x)$:[13]

$$\max_x EU(x, \theta) = v(x, \theta) - pf(x). \tag{2}$$

As $f(.)$ is a step function which is 'flat' between the different cutoffs x_i, the first-order condition for an *interior* optimum x^* is $\partial v(x^*;\theta)/\partial x = 0$. With the additive separable utility function, the interior option is independent of the enforcement parameters. x^* is thus equal to the driver's optimal speeding decision absent any penalty scheme. Our assumptions on $v(.)$ further imply that $x^* = x^*(\theta)$ is continuously increasing in θ.

While the interior solution does not depend on the penalty, the stepwise shape of $f(x)$ gives rise to possible *corner solutions*. Fig. 3 illustrates this point graphically, plotting x on the horizontal and expected utility (*EU*) on the vertical axis. The stepwise penalty scheme implies that the inverted-U shape of *EU* discontinuously drops at each cutoff x_i (in the graph: x_1 and x_2), as penalties increase at these speed level cutoffs ($\Delta_i > 0$). As illustrated in the left panel of Fig. 3, a notch does not necessarily imply a corner solution. Only if the *expected* notch, $p\Delta_i$, is sufficiently large (for a given driver θ), the driver's optimal speed corresponds to a corner solution at a cutoff. This case is depicted in the right panel of the figure.

4.1. Responses to notches

To study the impact of a penalty scheme with notches more formally, we follow the theoretical analysis in Kleven and Waseem (2013). Note first that, absent any penalties, the function $x^*(\theta)$ would simply map types of drivers into speed choices. The observed distribution of speed, $H(x)$, would be continuous. In the presence of notches, this will generally not be the case. This point is illustrated in Fig. 4, which plots expected utility for two types who face a notch at x_i. The interior solution for the driver of type $\underline{\theta}_i$ exactly corresponds to the cutoff, $x^*(\underline{\theta}_i) = x_i$, i.e., he would choose the cutoff x_i even absent any penalty scheme. Drivers with slightly higher θ are strictly better off when choosing their corner solution x_i rather than their interior solution $x^*(\theta) > x_i$ which would trigger a penalty $f_{i+1} > f_i$. In contrast, the driver with $\overline{\theta}_i$ is indifferent between his interior solution, $x^*(\overline{\theta}_i) > x_i$ and the corner solution at x_i. For the case depicted in Fig. 4, all types above $\underline{\theta}_i$ and below $\overline{\theta}_i$ strictly prefer the corner solution x_i over any speed $x > x_i$. The set of drivers who bunch at the cutoff x_i is thus given by the type interval $[\underline{\theta}_i, \overline{\theta}_i]$. The following Lemma generalizes this observation for a penalty scheme with more than one notch:[14]

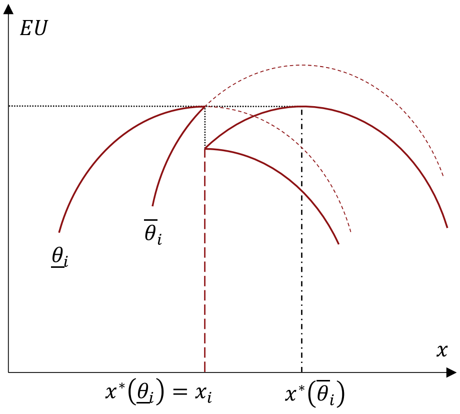


Fig. 4. Bunching at notch x_i. *Note:* The figure displays the mapping of speed x into expected utility for heterogeneous θ-types and one notch at the cutoff x_i. For the driver with type $\theta = \underline{\theta}_i$, the corner solution at the cutoff is identical to her interior solution. The driver with $\theta = \overline{\theta}_i$ is indifferent between her the interior solution $x^*(\overline{\theta}_i)$ and the corner solution at x_i. All drivers with $\underline{\theta}_i < \theta < \overline{\theta}_i$ will prefer the speed x_i over their interior optimum $x^*(\theta)$.

Lemma 1 . *Consider a notch $\Delta_i > 0$ at speed cutoff x_i that is used by a non-empty set of types. The probability mass of drivers speeding at x_i is given by $\Pi_i = G(\overline{\theta}_i) - G(\underline{\theta}_i)$, where $\underline{\theta}_i$ satisfies either*

$$x^*(\underline{\theta}_i) = x_i \tag{3}$$

$$\text{or} \quad \exists\, j < i\colon v(x_i, \underline{\theta}_i) - v(x_j, \underline{\theta}_i) = p(f_i - f_j), \tag{4}$$

and $\overline{\theta}_i$ satisfies either

$$\exists\, j > i\colon v(x^*(\overline{\theta}_i), \overline{\theta}_i) - v(x_i, \overline{\theta}_i) = p(f_j - f_i) \quad \text{and} \quad x_{j-1} < x^*(\overline{\theta}_i) < x_j, \tag{5}$$

$$\text{or} \quad \exists\, j > i\colon v(x_j, \overline{\theta}_i) - v(x_i, \overline{\theta}_i) = p(f_j - f_i). \tag{6}$$

Lemma 1 shows that there could be different characterizations of the boundaries of the interval $[\underline{\theta}_i, \overline{\theta}_i]$. In Fig. 4, the boundaries are characterized by Eqs. (3) and (5) (with $j = i + 1$). In the case of multiple notches, however, the lower bound could also be given by

[13] Risk aversion would not qualitatively affect our analysis as long as cross derivatives of Bernoulli utility functions with respect to the net benefits of driving and penalties are zero: In this case, the interior optimum x^* would remain unchanged; only the slopes of indifference curves for speed levels other than the optimum would be affected. The more risk averse a driver (and the larger the fine), the steeper the indifference curve. As a consequence, we would expect to observe *less* bunching for more risk averse drivers. Things would look different, however, if the aforementioned cross derivatives are non-zero: A driver's interior optimum x^* would then depend on her type *and* the size of the penalty.

[14] All proofs are provided in the Appendix.

Eq. (4). For this case, the type with $\underline{\theta}_i$ would be indifferent between a corner solution at x_i and a cutoff at a lower speed bracket, $x_j < x_i$. Similarly, the upper bound could be characterized by Eq. (6), which describes a type who is indifferent between a corner solution at x_i and a corner solution at a cutoff for a higher speed bracket, $x_j > x_i$.

The key implication from Lemma 1 is that notches may push drivers into corner solutions. Empirically, we should thus observe bunching of drivers at a speed level equal to a cutoff, $x = x_i$, and a sparsely populated (or even empty) range of speed levels slightly above a cutoff. The latter 'density holes' in $H(x)$ stem from drivers in the interval $[\underline{\theta}_i, \overline{\theta}_i]$ who would, in the absence of law enforcement (for $p = 0$), choose a speed in the range $x^*(\underline{\theta}_i) < x_i < x^*(\overline{\theta}_i)$. We will discuss the sensitivity of these predictions below.

4.2. Theoretical responses to the reform

To assess the impact of the 2009 reform on speeding, let us first consider a simple, hypothetical reform: an increase in f_h by a constant amount for all speed brackets $h > \ell$. Such a reform increases Δ_ℓ at cutoff x_ℓ but leaves all other notches unaffected. All speed levels $x > x_\ell$ become less attractive and drivers will choose a weakly slower speed. In terms of the distribution $H(x)$, some mass of drivers located above x_ℓ before the reform will be shifted towards lower speed levels.

Lemma 2 . *Consider a reform that increases one notch Δ_ℓ and leaves all other Δ_j, $j \neq \ell$ unchanged. Then every driver will drive weakly slower after the change.*

The hypothetical reform will also affect bunching. On the one hand, some types of drivers with an interior optimum $x^*(\theta) > x_\ell$ before the reform will start to bunch at cutoff x_ℓ (or a 'lower' cutoff x_l, $l < \ell$). Hence, bunching at x_ℓ (and lower cutoffs) tends to increase. On the other hand, driver types who were initially bunching at a cutoff x_h, $h > \ell$, will find it more attractive to drive slower. Bunching at x_h will then decrease.

Proposition 1 . *Consider a reform that increases one notch Δ_ℓ and leaves all other Δ_j, $j \neq \ell$ unchanged. The probability mass Π_i of drivers that bunch at a given cutoff x_i will then (i) weakly decrease if $i > \ell$, and (ii) weakly increase if $i \leq \ell$.*

In a nutshell, Proposition 1 shows that bunching at a given cutoff x_i increases in the size of a notch above this cutoff. Vice versa, bunching at x_i decreases if a notch at a lower speed cutoff increases. Obviously, the reform described in Table 1 differs from our hypothetical case. The 2009 reform was characterized by an increase of the notch at 20 km/h above the limit and of all notches at 30 km/h and above. The logic behind Proposition 1 thus predicts an increase in bunching at the 20 km/h cutoff (and at all lower cutoffs) after the reform. For all other cutoffs, however, the impact of the reform is ambiguous. To see this point, consider the 30 km/h cutoff, which experienced the largest increase of the notch. The larger notch at 30 km/h (and the increase of 'higher' notches) tends to induce more bunching at this cutoff, whereas the increase of the notch at 20 km/h works in the opposite direction. Without further assumptions, the overall effect on the bunching mass at the 30 km/h cutoff is therefore unclear. Independently of the bunching, however, our analysis suggests that the reform should result in weakly slower speed levels (see Lemma 2).

4.3. Discussion

The analysis from Section 4.1 suggests that we should expect bunching at cutoffs together with density holes in the speed range above a cutoff. There are, however, several arguments why this might not be borne out by the data. The most important arguments are based on the difficulties in targeting a specific cutoff x_i. First, there is substantial variation in the speed indicated by speedometers of different automobiles. Hence, a driver who observes a speed of, for instance, 130 km/h on the car's speedometer will most likely not drive $x = 130$ km/h.[15] This also means that cruise controls (which are fairly uncommon in Germany) do not necessarily facilitate the targeting of cutoffs.

Second, our notation indicates that we model the drivers' choice over a *penalty-relevant* speed x, i.e., the actual speed after applying the tolerance rule (see Section 2). Choosing a speed which corresponds to a cutoff x_i thus requires drivers to correctly compute the way in which the tolerance rule maps the measured speed s into the penalty-relevant speed x. While our survey suggests that every third driver exactly knows the tolerance rule, two thirds either over- or underestimate the rule's generosity. As a result there might be a significant amount of optimization errors — which are, presumably, more frequent than in the context of tax notches (Chetty, 2012; Kleven and Waseem, 2013). These errors will work against bunching and will diminish any density holes.

A further and important reason why we might not see much bunching is based on the possibility that drivers – particularly those who choose to drive well above the speed limit – underestimate the risk of a speed control. To see this recall that, *cet. par.*, bunching will be proportional to the *expected* notch, $p\Delta_i$. The lower the drivers' prior about the probability p, the more likely they are to choose their interior optima. If most speeders grossly underestimate the detection risk, we should therefore observe no or little bunching. However, if there is a second type of speeders, who drive above the limit despite a reasonably high prior about p, they will contribute to bunching. In combination, heterogeneous priors might therefore produce bunching without any pronounced density holes.

Note further that driving with very high speed (40 km/h above the limit and more; see Table 1) indicates that these drivers are willing to face temporary driving bans for one or several months (as long as they expect $p > 0$). For such drivers, the increment in monetary fines that occurs at, e.g., the 50 km/h cutoff might not be all too relevant. Put differently, from these speeders' perspectives, a local increase in fines by 100 Euro might be relatively small. Ultimately, this suggests that there might be little bunching at 'higher' speed cutoffs. Within our model we would obtain this prediction if the 'taste for speeding' (captured by the curvature of $v(x,\theta)$) would become more 'sharp' for high values of θ (as reflected in $\partial^3 v(x,\theta)/\partial x^2 \partial\theta$). We will return to this point below.

A last point worth discussing is the fact that our analysis – in contrast to the taxation literature (e.g., Chetty et al., 2011; Kleven and Waseem, 2013) – does *not* link the bunching mass to the elasticity of speeders with respect to the fine. There are two important reasons for not performing such an analysis. First, agents respond to two dimensions of policy: the penalty scheme (reflected in $f(x)$) and the detection probability p. While information on the scheme are available to drivers (just like tax rates and thresholds are observable for taxpayers), probabilities are largely unknown. In the context of heterogeneous priors one would then need very strong assumptions to identify the relevant elasticity.[16] Second, the type range of drivers who bunch in our context might be given by Eqs. (4) and (6) from Lemma 1 — a case which is neglected in tax studies, because tax thresholds are typically located in quite different (e.g., income) ranges. Our set-up, however, is characterized by many 'closely' located notches. Technically, we would need to identify both the highest and the lowest types of drivers

[15] The European Council Directive 75/443/EEC and §57(2) of the German *Straßenverkehrs-Zulassungsordnung* allows a tolerance in the speed displayed by speedometer of up to 13% *above* the true speed level (for the relevant speed range studied below). An indicated speed of 130 km/h could therefore correspond to an actual speed of only 115 km/h.

[16] Note further that the elasticities which derive from bunching estimates are sensitive to the specific functional form of preferences. While recent taxation research is characterized by an (implicit) consensus about the 'right' utility function, we are not aware of any consensus on drivers' preferences over speeding and monetary well-being.

bunching at a given notch, which implies one additional identification requirement.

5. Data

5.1. Highway data

To empirically evaluate speeding behavior, we use data from a highway police unit which is responsible for monitoring 575 km of Autobahn in the state of North Rhine-Westphalia. The police unit provided us with information on *all* tickets that emerged from speed controls on stretches of the Autobahn with a speed limit of 100 km/h during the period 01/2005 to 12/2006 and 01/2009 to 03/2010. All these speed controls were conducted using *mobile cameras* during periods of relatively low traffic, such that speeding is 'technically feasible' (i.e., there are no traffic jams). The data cover 154,970 speeding tickets with an overall amount of 10.85 m euros in fines. For each ticket we observe the penalty-relevant speed x in integer values (i.e., the outcome from applying the tolerance rule discussed in Section 2), the precise date, time and location of the speed measurement as well as the weather (sunny, cloudy, rainy) and street conditions (dry, damp, wet). For a subset of the tickets, we also observe the driver's gender and several digits of the vehicle's license plate. The latter information allows us to identify local drivers.[17]

Table 2, which provides summary statistics on our data, indicates that around 70% of the observations come from the pre-reform period. The data cover speeding tickets from all days of the week, with fewer tickets on weekends and slightly more on Wednesdays. More than 40% of the speeding offenses were recorded in the morning (8 am–12 am), around 25% in the evening (4 pm–8 pm) and less than 5% at night (8 pm–12 pm). For the sub-sample of tickets with slightly richer information we find that around 80% of speeders are male and roughly 20–30% are locals.

It is important to note that these data only include drivers that were speeding (and received a speeding ticket). Hence, the data do not allow us to observe the entire distribution of measured speed values. In fact, the police records reveal that not every speed measurement with $x > 100$ km/h resulted in a ticket. In 86% of all speed control sessions (covering 83% of all speeding tickets), the police only recorded and enforced speeding offenses with $x \geq 116$ km/h.[18] We therefore work with a *truncated distribution* of penalty-relevant speed measures. We can nevertheless derive some results on the general distribution (thus addressing extensive margin responses) by evaluating the data on speeders relative to the overall number of vehicles that is measured in each speed measurement session. In this vein, we can compute the fraction of vehicles speeding with more than 20 km/h above the limit, and so on. We will exploit this metric in Section 6.4.

Among the speeding tickets, the average speed x is 125.11 km/h, with a slightly lower speed in the post- as compared to the pre-reform period. Moreover, the average fine is 63.82 euro in the pre-reform sample. After the reform, this value increases by 35% to 86.29 euro. We will return to these observations in more detail below. The table further indicates several differences in the descriptives for the pre- and the post-reform sample. However, none of the differences in observables appears dramatic.[19]

5.2. Complementary data

As pointed out above, the Autobahn data only allow us to observe a truncated speed distribution, conditional on receiving a speeding ticket. In order to examine the full speed distribution, we gathered additional data from a large, permanently installed traffic control network in the state of Baden-Württemberg. The speed measures from this system are primarily used for the overall monitoring of trends in traffic flows and traffic density over time. The measures are *not* used for enforcing speed limits (i.e., a measure never triggers any legal consequences for drivers). The penalty scheme from Table 1 applies nonetheless.

These data cover single-lane roads. As in our main data, a speed limit of 100 km/h applies. We observe around 290,000 speed measures from 107 measurement points for the (post-reform) years 2010 until 2014. Comparing these data with the speeding ticket records, several points are worth stressing. First, the complementary data only cover cars (i.e., they do not cover trucks, larger buses or motorbikes). While the vehicle type is not recorded in the speeding tickets, it is very likely that tickets almost exclusively cover cars, too. Second, the speeding ticket records include information on the fine relevant speed x, whereas the complementary data directly provide integer values of the measured speed s.

Third, the main data stem from speed measures with mobile (laser) cameras, whereas the complementary data come from a fixed (visual and acoustic) system installed in, e.g., reflector posts that measure the 'average' speed based on the 'entry' and 'exit' of a vehicle in a given measurement range. Hence, even if we can retrieve the measured speed s from x (see below), the technical differences in the two measurement systems imply that one should be cautious in comparing the two sets of data. Fourth, the Autobahn data cover driving situations with two-lanes and relatively low traffic densities (see above). The complementary data, in contrast, cover speed measures for 24 h a day (for up to 14 days) which will certainly include periods of heavy traffic. This latter aspect, as well as the fact that the speed measure comes from an averaging of speed over a certain measurement stretch, suggest that the targeting of a certain speed level – and thus bunching – should be less pronounced in the complementary data-set.

6. Results

6.1. Descriptive evidence

We start out by examining whether the distribution of x among the speeding tickets provides any evidence for bunching at the cutoffs of the penalty scheme. Fig. 5 (a) illustrates the density distribution of the penalty-relevant speed x among the pooled sample of all speeding tickets with $x \geq 116$ relative to the total population of roughly 6 mio. measured drivers. The dashed green lines indicate the cutoffs from the penalty function (see Table 1). The density distribution is decreasing in the speed level and displays several major spikes. Two of these spikes are located right at cutoffs (120 and 125 km/h), and one is located slightly below a cutoff (129 km/h). The figure does not show any pronounced density holes 'to the right' of the spikes — a point that we will return to below. For speed levels with $x \leq 122$, we observe a lot of variation in the density distribution.[20] This makes it hard to evaluate the spike at the lowest cutoff (120 km/h). While it is more clear that there is no visible evidence for bunching at higher speed cutoffs (140, 150 and 160 km/h), the distribution shows several drops, some of which overlap with the cutoffs.

Recall that our data capture the penalty-relevant speed x *after* applying

[17] A driver is coded as local if the *Kreis* (county) indicated by the license plate corresponds to the 'home' or a neighboring Kreis of the location of the speed measurement.

[18] This practice is a response to the administrative costs of issuing and enforcing a ticket. These costs render speeding tickets with small fines economically unattractive. However, our data indicate that sometimes the police does enforce minor speeding violations (starting with 106 km/h).

[19] Examining the data at the level of measurement sessions (rather than for single speeding tickets), we do not find evidence on any radical change in the way when, where and for how long speed measures where conducted. The average duration of a measurement session, for instance, is 212 min in our pre- and 210 min in the post-reform data.

[20] We discussed this observation with the police unit which provided us with the data. While we could not identify any plausible explanation for the variation in the density for speed $x \leq 122$, we can reject the hypotheses that the variation is induced by different measurement techniques, rounding issues, or by overlapping enforcement thresholds.

Table 2
Summary statistics — speeding tickets.

	Pooled data		Pre-reform		Post-reform	
Variable	Mean	(Std. Dev.)	Mean	(Std. Dev.)	Mean	(Std. Dev.)
Speeding x	125.11	(9.20)	125.16	(9.04)	124.99	(9.60)
Monetary fine f	70.23	(49.72)	63.82	(39.01)	86.29	(67.02)
Enforcement limit: $x = 116$ km/h	0.83	(0.38)	0.82	(0.39)	0.86	(0.35)
Enforcement limit: $x = 121$ km/h	0.10	(0.30)	0.12	(0.33)	0.03	(0.18)
Male drivers[§]	0.83	(0.37)	0.84	(0.36)	0.80	(0.40)
Local drivers[§]	0.27	(0.44)	0.32	(0.47)	0.17	(0.37)
12:00 am–7:59 am	0.00	(0.02)	0.00	(0.01)	0.00	(0.03)
8:00 am–11:59 am	0.43	(0.50)	0.46	(0.50)	0.38	(0.48)
12:00 pm–3:59 pm	0.28	(0.45)	0.24	(0.43)	0.36	(0.48)
4:00 pm–7:59 pm	0.24	(0.43)	0.25	(0.43)	0.21	(0.40)
8:00 pm–11:59 pm	0.05	(0.22)	0.04	(0.20)	0.06	(0.24)
January	0.09	(0.28)	0.11	(0.31)	0.04	(0.19)
February	0.07	(0.26)	0.06	(0.24)	0.11	(0.31)
March	0.10	(0.30)	0.09	(0.29)	0.11	(0.32)
April	0.09	(0.28)	0.07	(0.26)	0.13	(0.33)
May	0.11	(0.32)	0.11	(0.31)	0.12	(0.32)
June	0.08	(0.27)	0.07	(0.25)	0.11	(0.31)
July	0.07	(0.26)	0.08	(0.27)	0.06	(0.24)
August	0.09	(0.28)	0.09	(0.29)	0.07	(0.26)
September	0.11	(0.31)	0.12	(0.33)	0.07	(0.25)
October	0.09	(0.29)	0.11	(0.31)	0.06	(0.24)
November	0.06	(0.23)	0.06	(0.24)	0.05	(0.23)
December	0.04	(0.20)	0.03	(0.17)	0.07	(0.25)
Monday	0.16	(0.37)	0.16	(0.37)	0.15	(0.36)
Tuesday	0.15	(0.36)	0.16	(0.37)	0.12	(0.32)
Wednesday	0.21	(0.40)	0.20	(0.40)	0.23	(0.42)
Thursday	0.15	(0.36)	0.15	(0.36)	0.16	(0.37)
Friday	0.14	(0.34)	0.13	(0.34)	0.15	(0.35)
Saturday	0.09	(0.28)	0.08	(0.27)	0.09	(0.29)
Sunday	0.11	(0.31)	0.11	(0.31)	0.11	(0.31)
Number of speeding tickets	154,970		110,721		44,249	
Number of speed control sessions	1139		843		296	

Notes: The table presents summary statistics – sample means and standard deviations in parenthesis – on the speeding tickets from the pooled, the pre- and the post-reform sample. The pre-reform data cover the period 01/2005 to 12/2006 plus 01/2009; post-reform: 02/2009 to 03/2010. The small share of observations in post-reform January is due to the fact that the reform was introduced on 1 February 2009. [§] indicates that the variable is only observed among a sub-sample of tickets.

the tolerance rule. Note further that the way in which the tolerance rule transforms the measured speed s into the penalty-relevant speed x mechanically produces a concentration of tickets at some values of x. In particular, all measures with $133 \leq s < 135$ [$166 \leq s < 168$] will be recorded with $x = 129$ [$x = 161$] in our data. Hence, the spike at $x = 129$ [and at $x = 161$] which is illustrated in Fig. 5 (a) might be a result of the tolerance rule's non-injective mapping of s into x.[21]

To account for this mechanical effect, we empirically 'revert' the mapping. For each value of x which maps one-to-one into s, we first assign the correct speed measured s (in integer values). Omitting the values for $s = \{133,134,166,167\}$ we then estimate a higher-order polynomial function that approximates the observed distribution of speed tickets over s. Based on the estimated distribution we finally assign the density mass from $x = \{129\}$ [$x = \{161\}$] to the speed levels $s = \{133,134\}$ [$s = \{166,167\}$]. The resulting distribution is presented in Fig. 5 (b). The figure shows that the massive spike at $x = 129$ from Fig. 5 (a) considerably shrinks once we account for the rounding rule. Nevertheless, the projected distribution includes a significant heap at $s = \{133,134\}$ (corresponding to $x = 129$).[22] Hence, the spike below the 130km/h cutoff is only partially due to the rounding rule.

Next, we turn to the speed distribution in our complementary data on single-lane roads (i.e., not from the Autobahn). Recall from above, that these data contain the measured speed s. The distribution, which is provided in Fig. 6, offers several additional insights: First, there is no evidence on bunching at the actual speed limit ($x = 100$ km/h corresponding to $s = 103$ km/h). Second, there is some evidence on bunching at the cutoff with 10 km/h and 15 km/h above the speed limit (corresponding to $s = 114$ and $s = 119$ km/h, respectively). Third, there is no visible evidence for bunching at higher cutoffs. We will return to these observations below.

6.2. Estimation approach

To estimate the bunching mass at a cutoff x_i, we start from the empirically observed mass of tickets within the range $[s(x_i) - \delta; s(x_i)]$, where $s(x_i)$ indicates a cutoff from the penalty scheme (in terms of measured speed s) and $\delta \geq 0$ defines the bunching area below the cutoff (in integer km/h values of measured speed). Following the literature, we then assess this mass of speeders relative to the expected mass from a counterfactual distribution for the hypothetical case without a notched penalty scheme.

To obtain the counterfactual, we approximate the speed ticket distribution from Fig. 5.b by a polynomial function. More specifically, we estimate

$$C_s = \sum_{q=0}^{\overline{q}} \beta_q \, s^q + \sum_{r=(s(x_i)-\delta)}^{s(x_i)} \gamma_r \cdot \mathbf{I}[s = r] + \epsilon_s \tag{7}$$

where C_s indicates the share of drivers measured with speed s, $\overline{q}$ defines

[21] To illustrate the problem, note that for any $133 \leq s < 135$ the tolerance rule – rounding down, subtracting 3% of the speed and rounding down again – will transform s into $x = 129$. For $135 \leq x \leq 165$, however, the tolerance rule is bijective, mapping *one* value of measured speed s into *one* observed value of x.

[22] Note that our approach assumes that the distribution of the observations from one speed level x among the two speed levels in s follows the estimated, 'smoothed' distribution. Hence, the projection ignores that rational drivers could anticipate the property of the tolerance rule and locate predominantly at the higher of the two speed levels s. If drivers indeed behave like this, the 'excess mass' would concentrate on $s = 134$. The following analysis will account for the fact that we cannot determine the precise speed measure for the two pairs of s.

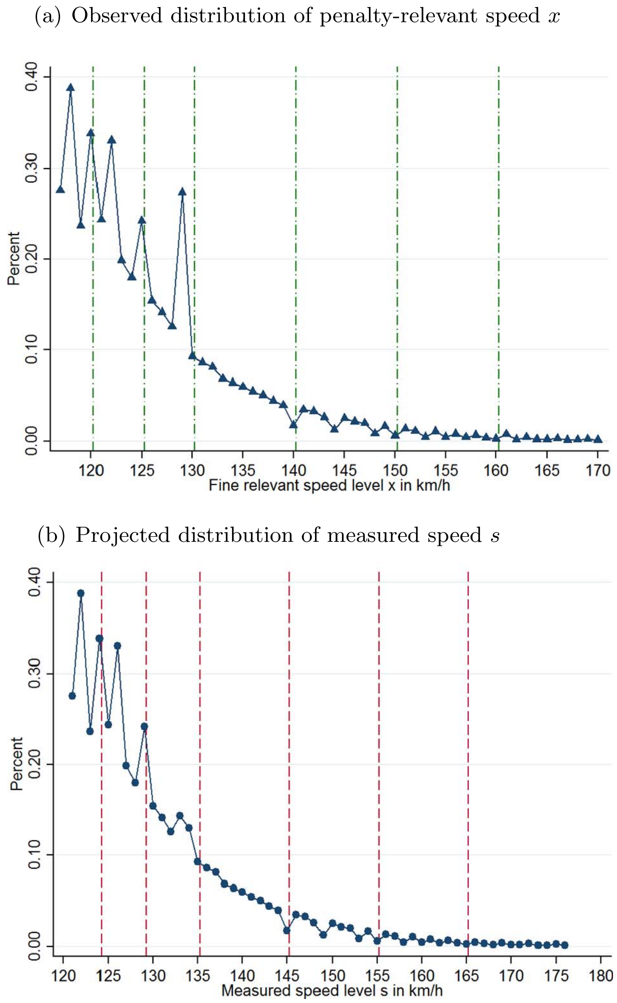


Fig. 5. Density distribution of speed. *Notes:* The figure illustrates the observed distribution of the penalty-relevant speed x (Panel a) as well as the projected distribution of measured speed s (*Panel b*) among all tickets (pooling data from the pre- and post-reform period). The vertical axes indicates the fraction of tickets observed for a given speed level, relative to the total number of speed measures. The horizontal axes capture the penalty-relevant speed x (Panel a) and the measured speed s (Panel b), respectively. The speed limit is 100 km/h. The dashed green lines indicate the cutoffs x_i from the penalty function (Panel a); the dashed red lines (Panel b) express these cutoffs in terms of the measured speed s. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the order of the polynomial function, and **I** is the indicator function. Based on the estimated β-coefficients (but excluding the γ-coefficients) we then predict $\hat{C}_s^{prox} = \sum_{q=0}^{\overline{q}} \hat{\beta}_q s^q$. This initial proxy for the counterfactual distribution neglects the excess mass of speeders from the range $[s(x_i) - \delta;s(x_i)]$ who would, in the absence of a notched penalty scheme, choose a speed level 'to the right' of the cutoff $s(x_i)$. To account for this fact, $\hat{C}_s^{prox}$ is inflated for speed values $s > s(x_i)$, up to the point where the counterfactual distribution of $\hat{C}_s$ satisfies $\sum \hat{C}_s = \sum C_s$ (i.e., when the empirical and the counterfactual distribution cover an equal number of speeding tickets). The bunching mass $\hat{b}_i$ in the speed range $[s(x_i) - \delta;s(x_i)]$ is then given by

$$\hat{b}_i = \sum_{s=(s(x_i)-\delta)}^{s(x_i)} \frac{C_s - \hat{C}_s}{\hat{C}_s/(1+\delta)}. \tag{8}$$

$\hat{b}_i$ indicates the excess mass, i.e., the difference between the observed and the predicted speed tickets with $s(x_i) - \delta \leq s \leq s(x_i)$ (in the numerator) relative to the average mass in the counterfactual distribution for this range (denominator). We estimate $\hat{b}_i$ together with boot-strapped standard errors using the iterative procedure from Chetty et al. (2011).

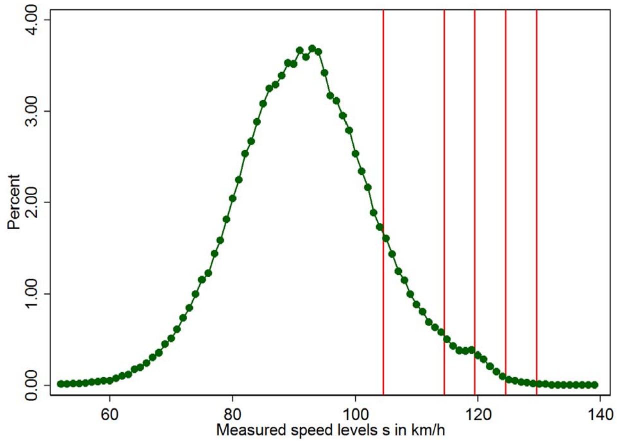


Fig. 6. Density distribution of measured speed s (complementary data). *Notes:* The figure illustrates the observed distribution of measured speed s among the complementary data ($N = 289, 125$; speed limit is 100 km/h). The vertical axes indicates the fraction of speed measures observed for a given speed level. The red lines indicate the speed limit as well as the cutoffs x_i from the penalty scheme in the s space, i.e., accounting for the tolerance rule. (The speed limit of $x = 100$ km/h is thus located at $s = 104$, etc.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Several aspects of our approach deserve a closer discussion. Note first that we base our estimates on the projected distribution of speed s rather than the distribution of penalty-relevant speed values x. By doing so, and by accordingly adjusting δ, we avoid potential problems with the rounding rule. Second, we will report bunching estimates that *locally* approximate the counterfactual distribution for the speed range around each cutoff. Our results remain qualitatively unaffected when we estimate one counterfactual for the full range of s. In this case we also included dummies for round numbers (see Kleven, 2016); our findings were again qualitatively robust. The same holds true for more simple approaches to quantify bunching (in the spirit of Saez, 2010).

Last, the estimation of Eq. (7) accounts for the observations from the bunching area $[s(x_i) - \delta;s(x_i)]$ but not for those in the range 'to the right' of a cutoff, with a potential missing mass in the distribution (compare Kleven and Waseem, 2013). This approach, which is closer to a 'kink'- rather than a notch-bunching analysis, is motivated by the fact that we face a high number of nearby notches with only few (integer valued) observations between two notches. This prevents us from jointly estimating a bunching and a missing-mass area (as in Kleven and Waseem, 2013). Moreover, we do not observe any pronounced density holes (see Fig. 5).

6.3. Bunching estimates

Fig. 7 (a) and (b) presents the results from bunching estimates for the main sample, focusing on the cutoffs with 25 and 30 km/h above the limit, respectively. For the moment, we pool the data for the pre- and post-reform period. Fig. 7 (a) shows sharp bunching right at the cutoff $s(x_i) = 129$. The estimated bunching coefficient for the range $s = \{128,129\}$ (i.e., at $x_i = 125$ with $\delta = 1$) indicates an economically and statistically significant excess mass of 36% (relative to the average counterfactual mass in that speed range). Fig. 7 (b), which presents the estimate for the cutoff at $s(x_i) = 135$, illustrates an excess mass which is located at least one km/h below the cutoff. To account for the fact that we cannot distinguish the measured speed for $s = 133$ and $s = 134$ (see Section 6.1), we estimate bunching in the broader range $s = \{133,134,135\}$ (i.e., we set $\delta = 2$ for $x_i = 130$).[23] Just like for the first cutoff, we obtain a significant bunching mass of 42%.

[23] With $\delta = 2$, the estimated coefficient is insensitive to how the projection allocates speed tickets from $x = 129$ to $s = 133$ and $s = 134$ (see Section 6.2).

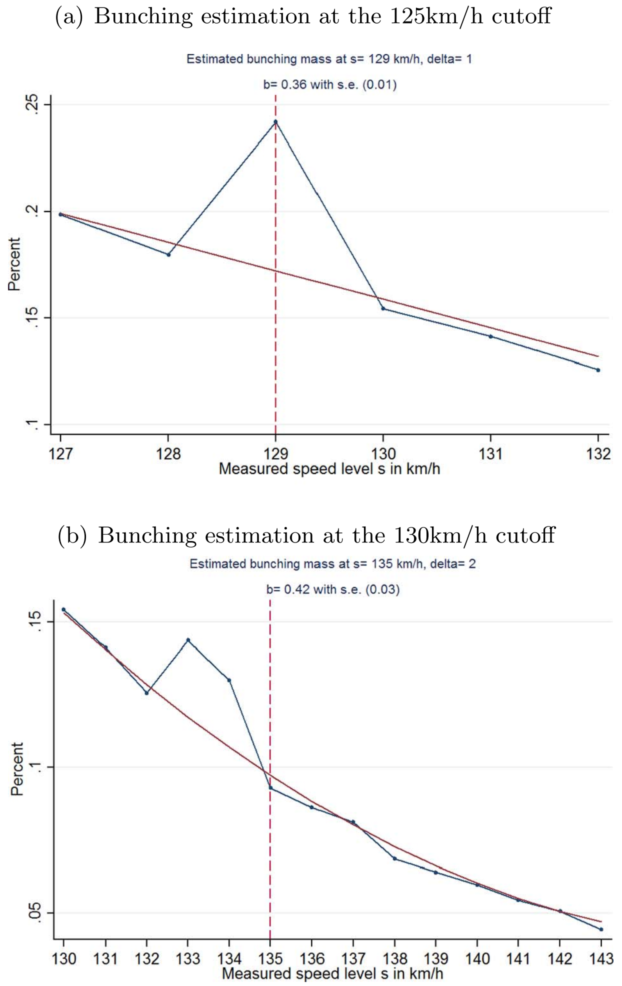


Fig. 7. Empirical and counterfactual distribution of speeding levels. *Note:* Empirical and counterfactual distribution of measured speeding for a speed limit of 100 km/h from pre- and post-reform data (pooled). The counterfactual distributions for graphs (a) and (b) are based on a linear (i.e., $\overline{q} = 1$) and a quadratic fit ($\overline{q} = 2$ for $s(x_i) = 135$), respectively. The horizontal axes indicates the empirical speed above the limit. The vertical axes indicates the percentage share of observations for each speed level (relative to all measured drivers). The red dashed vertical line in the *top* graph indicates the speed $s = 129$ km/h (corresponding to a penalty-relevant speed $x = 125$ km/h). The red dashed vertical line in the *bottom* graph indicates the speed of $s = 135$ km/h (corresponding to $x = 130$ km/h). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

What about bunching at the other cutoffs of the penalty scheme? Recall first that the majority of the speeding tickets are based on violations with $x \geq 116$. Hence, we cannot study bunching at the first two cutoffs, at 110 and 115 km/h. As discussed above, there is a sizeable spike together with a substantial amount of unexplained variation in the distribution of tickets around the cutoff at $x_i = 120$ km/h (compare Fig. 5). Our method to quantify bunching thus yields a positive but imprecisely estimated coefficient for this cutoff (see Table 3).

Consistent with the graphical evidence from Fig. 5, we do not detect any bunching for cutoffs at higher speed levels (see Table 3).[24] The null-results for these high speed cutoffs – which would correspond to an actually measured speed of $s = 145$, 155 and 165 km/h, respectively – are consistent with the arguments discussed above: Those who drive 40 km/h and more above the limit might either have a very sharp ‘taste for speeding’ (as captured by $\partial^3 v(x,\theta)/\partial x^2 \partial\theta > 0$) or they expect a very low detection risk *p*. In either case, they would be fairly insensitive to the notches in the fines. Recall further that penalties for these levels of speeding include significant driving bans, implying that the relative increase in the disutility associated with the changing *monetary* component of the penalty at these notch is rather small.

Table 3
Bunching estimates for different cutoffs.

Cutoff & bunching range	Pooled data	Pre-reform	Post-reform
$x_i = 120$ with $\delta = 1$	−0.04	−0.07	0.03
($s = \{123,124\}$)	(0.25)	(0.24)	(0.24)
$x_i = 125$ with $\delta = 1$	0.36	0.39	0.26
($s = \{128,129\}$)	(0.01)	(0.02)	(0.03)
$x_i = 130$ with $\delta = 2$	0.42	0.43	0.40
($s = \{133,134,135\}$)	(0.03)	(0.04)	(0.08)
$x_i = 140$ with $\delta = 1$	−0.54	−0.55	−0.51
($s = \{144,145\}$)	(0.15)	(0.13)	(0.19)
$x_i = 150$ with $\delta = 1$	−0.20	−0.30	0.06
($s = \{154,155\}$)	(0.56)	(0.57)	(0.62)
$x_i = 160$ with $\delta = 1$	−0.50	−0.57	−0.35
($s = \{164,165\}$)	(0.71)	(0.72)	(0.67)

Notes: The table displays the bunching estimates $\hat{b}_i$ for the cutoffs analyzed in Figs. 7 and A2 as well as for other cutoffs. x_i indicates the cutoff (in terms of penalty-relevant speed), s is the measured speed and δ captures the width of the bunching area.

Next, we apply these different steps of analysis to our complementary data. Computing local bunching estimates, we predict a modest (and insignificant) excess mass of 5% below the first cutoff (10 km/h above limit), and a slightly more pronounced (and statistically significant) 17% at the second cutoff (15 km/h above limit; both estimates for $\delta = 1$, as above; see Fig. A3 in the Appendix). Once more, we do not find any density holes on the right of the cutoffs. On the contrary, there seems to be an excess mass slightly above each cutoff. If we account for that and estimate the bunching mass for a speed range around the cutoff (in the spirit of a ‘kink’-estimator), we obtain significant excess masses of 12% for the first and 75% for the second cutoff (see Fig. A3). Hence, in line with the contextual and measurement differences between our main and the complementary data discussed in Section 5.2, we observe less sharp and more ‘fuzzy’ bunching in our complementary data. Finally, when we explore other possible bunching points, we detect no further evidence for bunching — neither at the speed limit itself nor at any of the higher cutoffs. The latter point, which was already noted while eyeballing Fig. 6, appears to be symmetric for the main and the complementary data.

We conducted several robustness checks and refinements (see Online Appendix, Section II). For our main data, the bunching estimators for the speed cutoffs at $x_i = 125$ and 130 km/h turn out to be highly robust to using alternative specifications (e.g., higher order polynomials) in the approximation of the counterfactual distribution. (For ‘higher cutoffs’, this is not the case; here we do observe much more variation in the estimated $\hat{b}_i$.) Splitting the sample for different hours of the day, different weekdays or seasons, we detect no significant differences in bunching (or stable null-results). We also split the sample according to different levels of traffic density (approximated by the number of measured cars per hour) and differentiated local and non-local drivers. Again, the data do not indicate any significant differences in bunching behavior.

We also estimated the probability of speed controls occurring at a given time (hour, weekday, month) and stretch of the Autobahn. This allows us to distinguish between ‘unlikely’ (or more surprising) and ‘likely’ (less surprising) speed controls. Similar to the results from other split-sample exercises, we only find minor, insignificant differences. What seems much more pronounced, however, is an overall adjustment in the drivers’ behavior: when speed controls are more likely to occur, we find considerably fewer violations of the speed limit in the first place. Comparing, for instance, the most- and least-surprising speed measurement sessions (the bottom- and top-tercile of predicted

[24] For the cutoff at $x_i = 140$ km/h, we obtain a significantly negative value for $\hat{b}_i$. We have no explanation for this puzzling finding (which is somewhat sensitive to the definition of the relevant local range).

Table 4
Propensity score matching of pre- and post-reform speeding.

Sample	Mean	(Std. Dev.)	1st quartile	Median	3rd quartile	*N*
Pre-reform	125.33	(9.16)	119	123	129	90,486
Post-reform	125.09	(9.69)	118	122	129	32,315

Notes: The table reports the comparison of pre- and post-reform speeding levels from a propensity matching exercise with the following confounders: time (hour of day, weekday, month), enforcement limits, location of speed control, street and weather conditions.

Table 5
Fraction of speeders relative to *all* measured cars.

Speed measure with…	Pre-reform	Post-reform
$x > 120$ km/h	0.030	0.023
	(0.028)	(0.021)
$x > 125$ km/h	0.019	0.014
	(0.020)	(0.012)
$x > 130$ km/h	0.011	0.008
	(0.013)	(0.008)
$x > 140$ km/h	0.004	0.003
	(0.006)	(0.003)
N (speed control sessions)	843	296

Notes: The table presents the fraction of speeding tickets with a speed x above different cutoffs from the penalty function, relative to all (speeding and non-speeding) cars measured per speed control session. Standard deviations are in parenthesis.

probability), the fraction of vehicles driving with more than 20 km/h above the speed limit (relative to all measured vehicles) drops from 1.41 to 1.09%. Hence, there seem to be pronounced extensive margin responses, suggesting that many drivers *do respond* to the variation in the objective detection risk. This also implies that the composition of the (self-)selected sample of speeders changes, which makes it hard to assess why bunching varies only modestly across measurement conditions.

To sum up, we find evidence for modest levels of bunching at some of the penalty scheme's notches. For the Autobahn data, there is bunching at the cutoffs 30, 25, and (less clearly) 20 km /h above the limit. The speed measures for single-lane roads contain weaker and more fuzzy bunching at cutoffs 10 and 15 km/h above the limit. In the light of the (i) difficulties in targeting the 'right' speed level (optimization errors), (ii) imperfect knowledge about the tolerance rule and the location of the cutoffs, and (iii) the impact from underestimating the detection risk, the evidence provides reasonable evidence on bunching. Several observations (and additional results discussed below) point to drivers' responsiveness and reject the case that drivers are unaware of the scheme (or 'fully schmeduling' in the sense of Liebman and Zeckhauser, 2004).

A striking difference to the literature on tax notches is the absence of density holes. In fact, the bunching observed for the complementary data looks more like the response to a kink rather than to a notch. As discussed above, this might be explained by optimization frictions and heterogeneous beliefs about p. On the one hand, drivers who anticipate a sufficiently high detection risk but nevertheless decide to speed might choose their optimal speed only among different cutoffs. For these drivers, the *expected* notches, $p\Delta_i$, would be large and a corner solution would dominate all interior solutions. Speed levels between two cutoffs x_i and x_{i+1} would only be observed due to optimization errors. On the other hand, drivers who believe that a speed control is unlikely to occur would choose their interior optima, which are smoothly distributed all over the speed range. The combination of drivers with heterogeneous beliefs could then produce some bunching without having any density holes in the distribution of speeding tickets. We will return to this discussion in Section 7.

6.4. *Actual responses to the reform*

Let us now turn to the impact of the reform. As we build on a pre/post-comparison, one should interpret the following results with caution. To start out, recall that the descriptive statistics from above revealed that, among the speeding tickets, the average speed x declined from 125.16 to 124.99 km/h (see Table 2). These numbers are clearly affected by other time-varying factors, i.e., when and under which conditions the speed was measured (see Section 5).[25]

To account for this issue, we only includes data for which the observable dates and circumstances of the post-reform measurement sessions have an equivalent in the pre-reform period. Based on these data we run a propensity score matching exercise to arrive at a pre- and post-reform sample that is comparable regarding the time (hour of day, weekday, month), the enforcement limit, as well as street and weather conditions. The results from this matching exercise are presented in Table 4. Similar to the basic descriptives, the results indicate that the reform is associated with a modest 0.23% decline in the penalty-relevant speed x, from 125.36 to 125.07 km/h.

It turns out, however, that Table 4 (which solely indicates that the average speed *among* speeding tickets is lower in the post-reform period) gives a misleading picture on the impact of the reform. This point becomes obvious once we analyze the change in the speed distribution beyond the selected sample of speeders. To do so, we computed the *share of speeding tickets* with a recorded speed x above a given cutoff x_i, *relative to the total number of vehicles* (speeding and non-speeding) which were measured during each speed control session. The results from this exercise are presented in Table 5.

The table reveals a pronounced shift in the speed distribution. Relative to all drives, the fraction of speeders with $x > 120$ km/h – i.e., cars driving in the speed range for which the reform increased the fines (see Table 1) – drop from 3.0 to 2.3%. Considering the cutoffs x_i at 125, 130 and 140 km/h, we observe a similarly strong decline of roughly 25% in the fraction of speeders who got ticketed with $x > x_i$. While these pre/post-reform differences may relate to other time-varying factors, the pattern from Table 1 is again confirmed by propensity score matching. Hence, the data are consistent with the prediction from Lemma 2: increasing the fines at $x_i = 120$ km/h (and 'higher' cutoffs) renders speeding in this range less attractive. We observe a pronounced shift in the speeding distribution: the fraction of speeding tickets with $x > 120$ km/h drops by 23%, alluding to non-trivial extensive margin responses. The simplicity and high salience of the penalty scheme might thereby contribute to this level of responsiveness (compare Abeler and Jäger, 2015).

In a next step, we study whether bunching at cutoffs changed between the pre- and the post-reform period. Graphical evidence suggests that bunching at the two cutoffs $x_i = 125$ and 130 km/h is equally observed in the pre- and the post-reform sample (see Fig. A1 in the Appendix). To assess the changes in the speed distribution, we first consider a simple estimation framework,

$$\text{Bunching}_j = \rho_0 + \sum_{\ell} \phi_\ell \, (\text{Post}_j \times I_j^{\ell}) + \sum_{\ell} \rho_\ell \, I_j^{\ell} + \mathbf{X}_j\kappa + \varepsilon_j, \tag{9}$$

where Bunching$_j$ is a dummy indicating whether a ticket j with speed x_j

[25] We do not detect any evidence suggesting that speed limit enforcement activities were massively increased after the reform. In fact, we observe a modest drop in the frequency of mobile speed control sessions (from 33.7 monthly sessions in our pre-reform sample to 24.9 monthly sessions for the year 2009). The average duration of a single session did not change much either (212 min in the pre- and 210 min in the post-reform period). The only noticeable change we observe is a less frequent use of fairly high enforcement thresholds (like 121 km/h, see Table 2) that is mirrored in a more frequent use slightly lower enforcement cutoffs (in the range between 100 and 115 km/h). The following analysis will account for these changes, which appear plausible given the incentive structures noted above (see fn. 18).

Table 6
Impact of reform on bunching.

	(1)	(2)	(3)	(4)	(5)	(6)
Reform × I^{120}	0.047***	0.050***	0.057***	0.047***	0.047***	0.054***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Reform × I^{125}	−0.014**	−0.022***	−0.023***	−0.020***	−0.027***	−0.027***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.008)
Reform × I^{130}	−0.013**	−0.024***	−0.026***	−0.017**	−0.027***	−0.026***
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.009)
Reform × I^{140}	−0.025*	−0.038***	−0.037**	−0.013	−0.024	−0.030
	(0.014)	(0.014)	(0.016)	(0.017)	(0.017)	(0.019)
Reform × I^{150}	0.013	−0.005	−0.014	0.023	0.007	0.008
	(0.025)	(0.025)	(0.028)	(0.027)	(0.027)	(0.030)
Reform × I^{160}	0.062	0.036	0.007	0.066	0.045	0.021
	(0.039)	(0.039)	(0.044)	(0.047)	(0.047)	(0.051)
Control variables:	No	Yes[a]	Yes[b]	No	Yes[a]	Yes[b]
N	154,970	154,970	128,644	154,970	154,970	128,644
R^2	0.665	0.668	0.668	0.571	0.574	0.573

Note: The table presents the outcome from LPM estimates of Eq. (9). All specifications include (non-interacted) cutoff specific dummies I_j^ℓ. In columns (2) and (5), we control for the year and the enforcement limit of the speed control session. Columns (3) and (6) add further control variables (for the weather conditions, location, month, day of the week and hour of the day). Columns (1)–(3) are based on $\delta = 2$, i.e., we set $I_j^\ell = 1$ if $x_\ell - 2 \leq x_j < x_\ell + 2$. Columns (4)–(6) employ $\underline{\delta} = 1$ and $\overline{\delta} = 2$ and thus set $I_j^\ell = 1$ if $x_\ell - 1 \leq x_j < x_\ell + 2$. The bunching dummies are adjusted accordingly. Robust standard errors are in parenthesis.

falls into the range at or slightly *below* a given cutoff x_ℓ, $x_\ell - \delta \leq x_j \leq x_\ell$. Post$_j$ indicates whether the ticket is from the post-reform period, I_j^ℓ captures if a speed ticket with x_j is located around a given cutoff, $x_\ell - \delta \leq x_j < x_\ell + \delta$, and $\mathbf{X}_j$ is a vector of control variables (including dummies for the hour, day, month, as well as street and weather conditions during the speed measurement).

For each cutoff ℓ, the coefficients ρ_ℓ then captures the fraction of tickets (among those in $x_\ell - \delta \leq x_j < x_\ell + \delta$) that are located at or slightly below x_ℓ. Hence, the ρ-coefficients will not capture bunching; they solely reflect the local slope of the (pre-reform) distribution around each cutoff (as captured in Fig. A1). The coefficients of interest are the ϕ's, which indicate how the fraction of tickets at or slightly below a cutoff changed after the reform. Linear probability model estimates of the ϕ's from Eq. (9) are presented in Table 6.[26]

Consistent with our theoretical prediction on the impact of the reform, we observe an increase in the mass of tickets at or slightly below the cutoff at 120 km/h. The estimates from Table 6 suggest that after the reform there is a 5 percentage point higher chance of seeing a ticket with a penalty-relevant speed just below 120 km/h in the data. While our theoretical framework does not offer any clear predictions regarding the reform's impact on bunching at other notches (see Section 4.2), it is interesting to note that the estimates point to a decline in the frequency of tickets below the cutoffs at 125 and 130 km/h — the two cases for which we found strong and robust bunching evidence above. There also seems to be a decline in the mass of tickets below the cutoff at 140 km/h, however, the estimate is sensitive to the precise specification and not robust when we vary δ. For the other cutoffs, the regression analysis does not indicate any significant impact of the reform.

To further assess the change in bunching, we also estimated the coefficient $\widehat{b}_i$ from Eq. (8) for the pre- and post-reform period. The results, which are presented in columns 2 and 3 of Table 3, capture again an increase in the mass of tickets right at the 120 km/h cutoff.[27] For the 125 and 130 km/h cutoffs, we observe a decline in bunching. For the former cutoff, the estimated excess mass drops from 39 to 26%; for the latter cutoff we estimate a more modest decline, from 43 to 40% (see Table 3 and Fig. A2 in the Appendix).

To wrap up, the second part of our analysis points to a non-trivial impact of the 2009 reform, which considerably increased all notches starting with 20 km/h above the speed limit (see Table 1). In line with theoretical predictions, we observe a 25% drop in the fraction of drivers speeding with 120 km/h or above. At the same time, the fraction of speeding tickets slightly below the 120 km/h cutoff increases in the post-reform period.[28]

7. Welfare

From a welfare perspective, one might argue that a notched penalty scheme might be inferior to the 'correct' Pigouvian correction mechanism. The latter would account for the marginal externalities from speeding (accident risk, air and noise pollution, etc.) as a continuous function of speed. Hence, a notched scheme is, at best, only a rough approximation to a Pigouvian scheme (Sallee and Slemrod, 2012). The simplicity of the stepwise scheme, however, might raise drivers' awareness and attention, thus contributing to a higher *salience* of and responsiveness to the penalty system (Abeler and Jäger, 2015). Different from other tax applications (Chetty et al., 2009; Finkelstein, 2009), this might be good from a social point of view.[29] Depending on this latter 'salience effect', the notched scheme could therefore dominate a more complex (but less salient), smooth penalty function. To examines this possible welfare advantage in more detail, we extend our formal analysis and impose additional structure. Building on our empirical findings, we then assess the tradeoff between the notched and a hypothetical, smooth penalty scheme.

[26] Estimates using non-linear models, which are available from the authors, yield almost identical results.

[27] Note that the bunching estimates use $\delta = 1$. If we set $\delta = 2$, as in the LPM, the bunching estimates show a more pronounced increase in the excess mass at the 120 km/h cutoff from 9 to 36%.

[28] Consistent with this pronounced shift in the speed distribution, we note a drop in aggregated accident statistics after the reform, too. Comparing 12 or 24 months before and after the reform (using either 12-month-differentiated estimates or estimates using month fixed effects with or without different pre- and post-reform trends), we find a decline in accidents for the post-reform period. (Estimates based on data obtained from *Deutsches Statistisches Bundesamt, Fachserie 8/7, Verkehr* (monthly data for 2007–2012).) A more careful analysis that aims at identifying if this inter-temporal variation is associate with a truly causal impact of the reform, is beyond this paper's scope.

[29] In the case of a distortive tax, a lack of salience implies a lower responsiveness of taxpayers and thus a lower deadweight loss. By contrast, the purpose of an externality correcting penalty scheme is to induce a change in drivers' behavior, such that higher salience is beneficial.

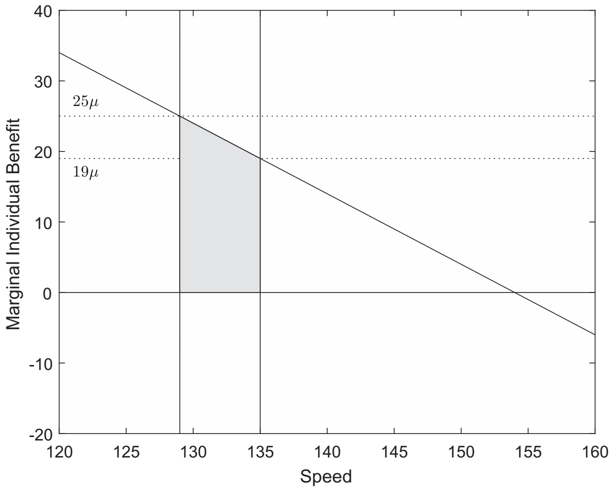


Fig. 8. Marginal benefits from speeding. *Note:* Using the functional form introduced in Eq. (11), the figure illustrates the marginal benefits of speeding as as a function of s for driver type $\hat{\theta}$ with an unconstrained optimum $s(x^*(\hat{\theta})) = 154$ km/h.

7.1. Pigou, salience and welfare

In the following, we have to account for the distinction between actual (s) and penalty relevant speeds (x), as discussed above. Given that our empirical analysis mainly builds on s (which is also the relevant speed for a driver's utility and welfare), we carry out our analysis in terms of this metric. We start out by assuming that the externality caused by speeding is proportionate to the speed. Let the welfare impact from a ride of a θ-driver be given by

$$W(s, \theta) = v(s, \theta) - e \max\{s - \underline{s}, 0\}, \tag{10}$$

where e is the marginal externality and $\underline{s}$ is a 'safe' speed, below which no externality occurs.[30] We assume that the actual speed limit reflects this save speed. The first-order condition for type θ's welfare maximizing speed $s^e(\theta)$ (provided it is above $\underline{s}$) is then $\partial v(s^e(\theta),\theta)/\partial s = e$. If limited attention and salience effects played no role, the welfare maximizing smooth penalty scheme would then be given by $f^s(s) = e^{'} \max\{s - \underline{s}, 0\}$ with $e^{'} := \frac{e}{p}$ (for $p > 0$).[31] As we consider a constant marginal externality, this scheme is fairly simple. Conceptually, however, one could easily replace it with a more general (and more complex) externality function $E(s)$.

Following Sallee and Slemrod (2012), we approximate the marginal individual net benefit from speeding ($\partial v(.)/\partial s$) to be linear in s. In particular, we assume that

$$v(s, \theta) = v - \frac{1}{2}\mu(\theta - s)^2. \tag{11}$$

Without any enforcement ($p = 0$), this implies $s^*(\theta) = \theta$. For $p > 0$, a driver facing the smooth, Pigouvian scheme $f^s(s)$ from above, will choose the welfare maximizing speed, $s^e(\theta)$. By deviating from the unrestricted, individually optimal speed, $s^*(\theta)$, the driver's disutility then equals $\frac{1}{2}e(s^*(\theta) - s^e(\theta))$. At the same time, the externality imposed by the driver shrinks by $e(s^*(\theta) - s^e(\theta))$, thus reflecting the overall welfare gain from moving from $s^*(\theta)$ to $s^e(\theta)$.

To compare the smooth, Pigouvian scheme with the notched penalties from Section 4, we will assume that the marginal penalty e' corresponds to the average notch-size (i.e., the average increase in penalties at the different cutoffs) smoothed over the relevant range of a given speed bracket (i.e., $e^{'} \approx \frac{1}{I}\sum_i \Delta_i/(s_i - s_{i-1})$, where I is the number of notches; see Section 4). Intuitively speaking, this is saying that the notched scheme 'approximates' the slope of the smooth scheme.

Let us now turn to the salience aspect. Recall that we empirically observe many drivers with a speed that is slightly above a speed level cutoff (i.e., choices in strictly dominated speed ranges; compare Fig. 3). Accounting for this fact, we assume that a fraction λ^n of drivers is inattentive and acts as if $p = 0$ under the notched scheme. These drivers do not bunch but myopically choose $s = s^*(\theta) = \theta$. In the following, we will refer to them as M-drivers. Vice versa, a share of $1 - \lambda^n$ rational R-drivers optimize as described in Section 4.

For the smooth penalty scheme, the fraction of M-drivers is denoted by λ^s. Allowing for a lower salience of the smooth scheme, we consider $\lambda := \lambda^s - \lambda^n \geq 0$; the fraction of M-drivers might be higher under the smooth as compared to the notched scheme. We will now turn back to our data to ask how large the 'salience gap' λ (i.e., the difference in the share of M-drivers across both schemes) could at most be, such that the welfare costs from the lower salience of the smooth scheme would be still outweighed by the gains from providing the correct Pigouvian incentives.

7.2. Smooth versus notched scheme

Let us first approximate the speed choices of drivers absent any penalty scheme. In doing so, we make use of our bunching evidence. Recall that we observe bunching at two cutoffs, $x = 125$ and $x = 130$ (corresponding to the driving speeds $s = 129$ and $s = 135$). The speed levels immediately above these cutoffs are strictly dominated. Hence, the drivers that we observe at these dominated speed levels must be M-drivers. The upper boundary of this dominated area would then be marked by a hike in observed drivers caused by R-drivers for whom that speed level is the interior optimum $x^*(\theta)$. However, we do not observe such a hike. One explanation that is consistent with this observation, is that the entire speed range above $x = 125$ is, except for the cutoff at $x = 130$, dominated. In other words, among those observed with $x \geq 125$, all R-drivers bunch at one of the two cutoffs $x = 125$ and $x = 130$. In order to approximate R-drivers' speed choices $s^*(\theta)$ absent a penalty scheme, we have to find the R-driver of type $\hat{\theta}$ who would be indifferent between these two cutoffs.

To this end, we assume that the types (and therefore the interior optima) are distributed in exactly the same way among R- and M-drivers. This implies that (i) the fraction of M-drivers with $128 \leq s \leq s^*(\hat{\theta})$ among M-drivers with $s \geq 128$ is equal to (ii) the fraction of R-drivers bunching at or below $s = 129$ relative to all R-drivers with $s \geq 128$.[32] Given our assumption that all R-drivers are bunching at the two cutoffs, this latter fraction is then given by

$$\frac{\sum_{s=128}^{129}(C_s - \hat{C}_s)}{\sum_{s=128}^{129}(C_s - \hat{C}_s) + \sum_{s=133}^{135}(C_s - \hat{C}_s)} \tag{12}$$

(where we account for the bunching range δ used in Section 6). Based on the observed distribution C_s and $\hat{C}_s$ from our estimates for Eq. (7) from Section 6.3, one can easily compute this fraction. Our data suggests that 19.25% of all drivers with $s \geq 128$ are R-drivers who bunch at cutoff $x = 125$ (the numerator in Eq. (12)), and 34.42% off all drivers who drive at $s \geq 128$ are R-drivers in total (the denominator in Eq. (12)). Based on this, we obtain $\lambda^n = 1 - 0.3442 = 0.6558$. Note that this high fraction of M-drivers – among those driving $s \geq 128$ – reflects our strong assumption that only those who bunch are R-driver. Neglecting optimization errors or rational drivers with very 'sharp' preferences (in particular, for very high θs and speed levels), our approach thus tends

[30] Of course, this is a gross simplification. In reality, *any* ride will generate an external effect.

[31] We treat p as exogenously given. Moreover, we neglect costs associated with providing a certain level of p and consider penalties as welfare neutral transfers.

[32] The latter point reflects that – among R-drivers with $s \geq 128$ and under our assumptions from above – all types with $\theta < \hat{\theta}$ would bunch at the lower, whereas all those with $\theta \geq \hat{\theta}$ would bunch at the higher cutoff. Note further, that we consider the speed range $s \geq 128$ to reflect the estimates from Section 6 which use $\delta = 1$ (at $x = 125$).

to overestimate λ^n.

Putting together these numbers, the fraction from Eq. (12) equals 55.93%. When we now examine at which point the mass below the estimated polynomial Eq. (7) for the range above $s = 128$ equals 55.93%, we arrive at the cutoff speed $s^*(\hat{\theta}) = 154$ km/h. This now approximates the speed, at which an R-driver with $\hat{\theta}$ – who would be indifferent between both cutoffs from above – would drive absent any penalty scheme.

In a second step, we now consider a ‘comparable’ smooth linear penalty scheme (comparable, in the sense of $e' \approx \Delta_i/(s_i - s_{i-1})$ as explained in Section 7.1). Building on the finding from above, we focus on a type $\hat{\theta}$ driver with an unconstrained optimum $s^*(\hat{\theta}) = 154$ km/h. Using the structure imposed in Eq. (11), Fig. 8 plots this type $\hat{\theta}$'s marginal benefits as a function of s (absent any penalties). The type's utility loss from driving at $s = 129$ instead of $s = 135$ km/h is given by the shaded area under the line between $s_1 = 129$ and $s_2 = 135$. The size of this area equals $(135 - 129)\left(\frac{1}{2}6\mu + 19\mu\right) = 132\mu$. The difference in the expected (smooth) penalties between the two speed levels would be $e(135\text{–}129)$ (where, by definition, $pe' = e$). Hence, under the linear scheme, a type $\hat{\theta}$ would still be indifferent between $s_1 = 129$ and $s_2 = 135$ if $e = 22\mu$. Substituting this level for e in the first-order condition $\mu(\hat{\theta} - s^e(\hat{\theta})) = e$, we arrive at the interior solution for this type's optimal speed under the smooth linear penalty scheme, which is $s^e(\hat{\theta}) = s^*(\hat{\theta}) - 22 = 132$ km/h.

In a third step, we can now predict the effect of replacing the observed notched with a smooth but less salient penalty scheme on speeding choices of types θ with $s^*(\theta) \geq 128$. The overall (non-) response of drivers' behavior to this hypothetical reform can be decomposed for three different groups. (1) A share of drivers λ will become inattentive, i.e., they turn from R- into M-drivers. These drivers will switch from bunching at one of the notches ($x = 125$ or $x = 130$) to their unconstrained interior optima, $s^*(\theta)$. (2) The remaining R-drivers, who account for a share $1 - \lambda^s$ of all drivers, will re-optimize to the Pigouvian scheme, choosing a welfare optimal speed level, $s^e(\theta)$. Finally, (3) all M-drivers under the notched scheme will remain M-drivers under the smooth scheme, again choosing $s^*(\theta)$. To examine the welfare implications from the first two groups' responses, one has to note that our assumptions regarding the linear marginal benefit of speeding and the constant externality imply that any change in speed from s_n to s_s (indicating the speed choice under the notched and the smooth scheme) has a welfare effect of

$$\begin{aligned}\Delta W &= \frac{\mu}{2}[(s^*(\theta) - s_n)^2 - (s^*(\theta) - s_s)^2] - e(s_s - s_n) \\ &= \mu\left[s^*(\theta) - \frac{s_n + s_s}{2} - 22\right](s_s - s_n).\end{aligned} \tag{13}$$

Let us first consider the share $1 - \lambda^s$ of R-drivers that rationally optimize under both schemes (group 2 from above). From above we know that, under the smooth scheme, $s_s = s^*(\theta) - 22$, which implies $\Delta W_2 = \mu\frac{(s_s - s_n)^2}{2}$. We use the estimated polynomial of the counterfactual distribution Eq. (7) to assign R-drivers to $s^*(\theta)$ and, thus, to $s_s = s^*(\theta) - 22$.[33] We weight the welfare gains ΔW_2 with the relevant frequencies and add them up. This exercise indicates that the size of the cumulative welfare increase for this group is 84.66μ. Note that every single driver from this group causes a modest welfare gain, as the Pigouvian scheme induces them to drive at the welfare optimal speed.

Next, consider the share λ of former R-drivers that turn inattentive and fall back to their interior optimum, $s_s = s^*(\theta)$, under the smooth scheme (group 1 from above). The welfare implications of their response is given by $\Delta W_1 = \mu\left[s^*(\theta) - \frac{s_n + s_s}{2} - 22\right](s_s - s_n) = \mu\left(\frac{s_s - s_n}{2} - 22\right)(s_s - s_n)$. Again, we calculate the welfare changes for all these types θ of R-drivers, weight them with the same frequencies derived from the estimated polynomials as above, and sum them up. This yields a welfare loss of 157.34μ for this group. As speeding choices of stable M-drivers remain unaffected, we thus arrive at the overall welfare change from the hypothetical change in the penalty scheme:

$$\Delta W = \mu[84.66(1 - \lambda^s) - 157.34\lambda] = \mu[84.66(1 - \lambda^s) - 157.34(\lambda^s - \lambda^n)]. \tag{14}$$

The overall welfare from replacing the notched with a less salient smooth scheme is non-negative if and only if $\lambda^s \leq 0.35 + 0.65\lambda^n = 0.78$, where the last step uses the result $\lambda^n = 0.66$ derived above. As compared to a notched scheme, an ‘equivalent’ smooth penalty scheme would thus be welfare enhancing if at most $\lambda = 12$ percentage points of drivers would turn inattentive under the less salient scheme (implying $\lambda^s \leq 66 + 12 = 78\%$). As long as the salience gap λ is smaller, a more complex Pigouvian scheme could be welfare enhancing.

7.3. Caveats and discussion

It is important to emphasize that our analysis is based on several strong assumptions and should be seen as a back-of-the-envelope calculation. We have already noted above that our approach (which treats many potential R- as M-drivers) will most likely yield an upper bound for λ^n. From Eq. (14) then follows, that the upper bound $\lambda = 12$ pp derived above might be fairly conservative. (As $\lambda^n \rightarrow 0$, the bound would approach $\lambda = 35$ pp.)

At the same time, however, one might argue that a truly Pigouvian scheme would be *extremely* complex, e.g, accounting for the fact that externalities vary not only with speed but also with driving conditions (weather, traffic density, etc.). For such a scheme, the ‘salience gap’ could be enormous, potentially rendering first-best correction of external effects infeasible (at least, as long as advanced automated driving systems do not yet fully take over speed choices). Hence, even a more complex but smooth scheme (as, e.g., the penalty function used in The Netherlands) would offer an imperfect correction of externalities. As a consequence, the welfare gains from replacing the notched scheme (ΔW_2) might be smaller, which would *cet.par.* impose even more tighter limits on λ.

Conceptually, our analysis omits a further point that should work in favor of the notched system: Bunching tends to (locally) reduce the variance in speed. A lower variance, in turn, could directly contribute to a reduction in the accident risk (Lave, 1985), which would entail positive welfare effects associated with the notches.

8. Conclusion

This paper studies drivers' knowledge of and responses to a notched penalty scheme for speeders in Germany. We provide survey evidence suggesting that many drivers have a good knowledge of the scheme's stepwise shape with its discontinuous jumps in penalties at certain speed cutoffs. The survey also documents heterogeneity across drivers, which – in combination with other optimization frictions – limits the scope for sharp bunching.

Exploiting micro-data from more than 150,000 speeding tickets from the German Autobahn and 290,000 speed measurements from traffic flow systems on other roads, we nevertheless find evidence on bunching at some of the penalty scheme's cutoffs. However, the bunching mass is fairly small. Moreover, at the speed limit itself and for notches at higher speed levels, we do not detect any excess mass in the distributions. The latter observation is consistent with the interpretation that excessive speeders might have

[33] For instance, for the lowest type R-driver bunching at $s = 129$, the interior optimum corresponds to this speed level, such that $s_s = 129 - 22 = 107$. This yields a welfare gain of $\Delta W_2 = \mu\frac{(s_S - s_n)^2}{2} = 242\mu$. Maintaining our assumption that θ types are identically distributed under R-types as under M-types, this is true for 2.30% of all those driving at $s \geq 128$. Similarly, we know that the highest type of R-driver bunching at $x = 125$ has an interior optimum at $s = 154$; he would choose $s_s = 132$ under the smooth scheme. This is true for 1.95% of those R-drivers driving at $s \geq 128$.

underestimated the risk of a speed control.

This point also highlights one major difference between our analysis and the bunching studies in the taxation literature. In our context, agents respond to *expected notches* which are jointly shaped by two policy parameters: the (shape of the) penalty function and the detection risk. With heterogeneous priors about the latter risk, one cannot directly translate the bunching mass from these notches into a straightforward measure of behavioral elasticities.

The fact that some drivers' speed choices seem suboptimal even under a salient, notched penalty scheme reinforces the concern that the responsiveness under a more complex, smooth scheme – that correctly accounts for externalities of speeding – might be even lower. To address this idea, we conducted a welfare analysis that models behavioral responses to notches as stemming from both salience and price responses. Based on our empirical results, we derive a bound for the 'salience gap' between a notched and a more complex but less salient Pigouvian scheme. If this salience gap – the fraction of drivers that turn inattentive under the more complex scheme – is below 12 percentage points, the welfare gains from applying a smooth Pigouvian scheme dominate the benefits associated with the enhanced salience of the notched scheme. Empirically evaluating the salience and behavioral consequences from introducing a smooth penalty scheme – as it is used, for instance, in The Netherlands – is up to future research.

Appendix A. Appendix

A.1. figures

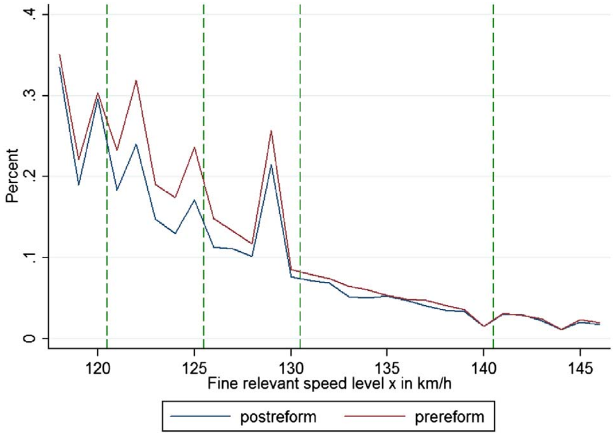
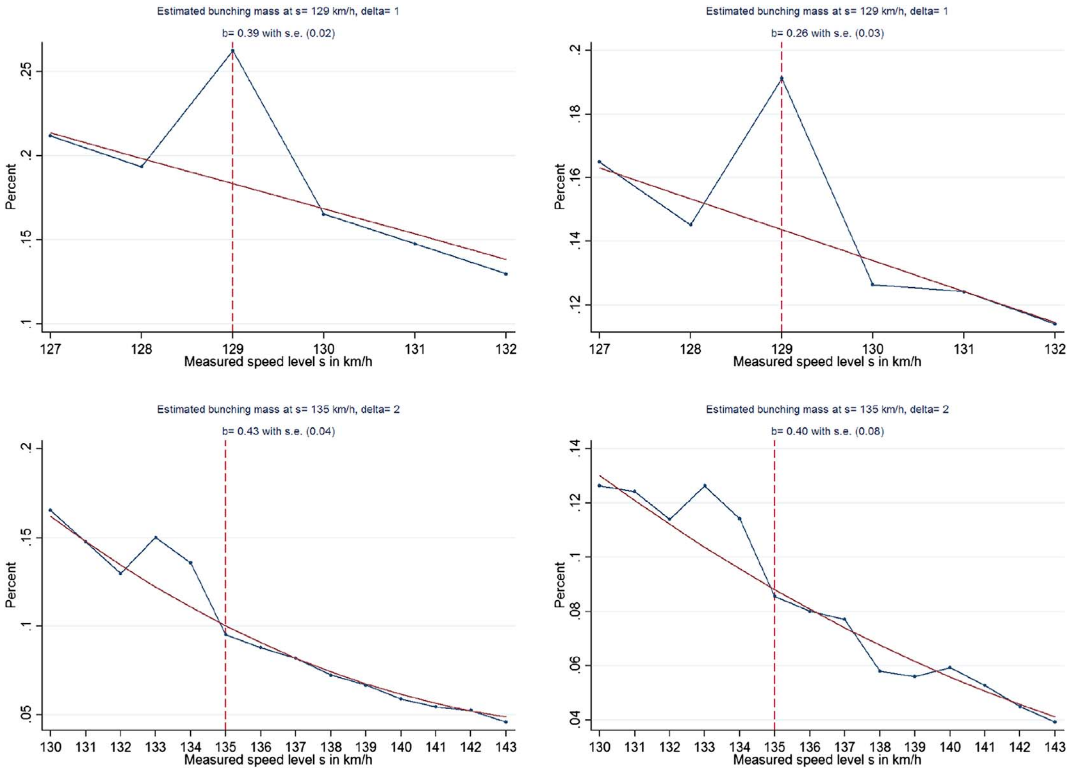


Fig. A1. Pre- and post-reform distribution of penalty-relevant speeding levels. *Note:* The figure illustrates the distribution of penalty-relevant speeding levels among speeding tickets from the pre- and post-reform period. The horizontal axis indicates the speed above the limit. The vertical axis indicates the percentage share of observations for each speed level. The green dashed vertical lines indicate the cutoffs at the respective speed levels. The number of observations for pre-reform [post-reform] period is 89,520 [34,356] in the displayed speed range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)


Estimated bunching mass at s= 129 km/h, delta= 1

b= 0.39 with s.e. (0.02)

Estimated bunching mass at s= 129 km/h, delta= 1

b= 0.26 with s.e. (0.03)

Estimated bunching mass at s= 135 km/h, delta= 2

b= 0.43 with s.e. (0.04)

Estimated bunching mass at s= 135 km/h, delta= 2

b= 0.40 with s.e. (0.08)

Fig. A2. Empirical and counterfactual distribution of speeding levels: pre- and post-reform period. *Note:* The two panels on the left display the empirical and counterfactual distribution of measured speed for the *pre*-period. The panels on the right show the distributions for the *post*-reform period. The top panels compares the distribution around the cutoff $x_i = 125$ (corresponding to a measured speed $s = 129$ km/h), the panels at the bottom consider the cutoff $x_i = 130$ (corresponding to $s = 135$ km/h). The horizontal axes indicate the measured speed *s*. The vertical axes show the share of observations for each speed level in percent. The blue line captures the empirical distribution and the red curve shows the estimated counterfactual distribution. The counterfactual in the top [bottom] panels are estimated with a linear [quadratic] slope. The dashed vertical lines indicate the cutoffs in terms of measured speed *s*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

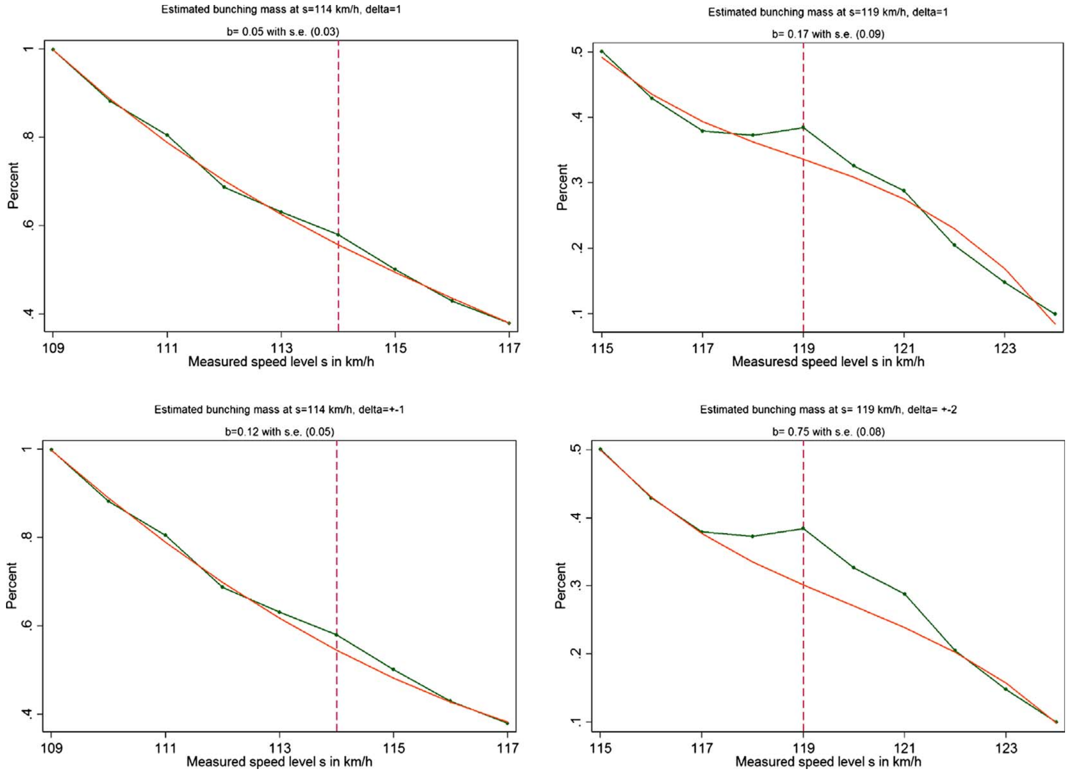


Fig. A3. Empirical and counterfactual distribution of speeding levels (complementary data-set). *Note:* The figures present the empirical and different estimates for the counterfactual distribution of speed s in the complementary data (speed limit of 100 km/h). The counterfactual estimates are all based on a cubic fit. The two graphs of the first row consider $\delta = 1$ (i.e., bunching at the cutoff and one unit below). Accounting for the actual allocation of observations, the graph on the bottom left and right compute the bunching mass for $\delta = \pm 1$ and $\delta = \pm 2$ (in the spirit of a conventional 'kink'-estimator), respectively. Horizontal axes indicate the empirical speed s. Vertical axes indicate the percentage share of observations for each speed level (relative to all measured drivers). The dashed, red vertical lines indicate the speed $s = 114$ and $s = 119$ km/h (corresponding to a penalty-relevant speed $x = 110$ and $x = 115$ km/h), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2017.11.006.

References

Abeler, J., Jäger, S., 2015. Complex tax incentives — an experimental investigation. Am. Econ. J. Econ. Pol. 7 (3), 1–28.

Altmann, S., Traxler, C., Weinschenk, P., 2017. Deadlines and Cognitive Limitations. In: IZA Discussion Paper No. 11129.

Bastani, S., Selin, H., 2014. Bunching and non-bunching at kink points of the Swedish tax schedule. J. Public Econ. 109, 36–49.

Blinder, A., Rosen, H.S., 1985. Notches. Am. Econ. Rev. 75, 736–747.

Bourgeon, J.-M., Picard, P., 2007. Point-record driving licence and road safety: an economic approach. J. Public Econ. 91 (1–2), 235–258.

Castillo-Manzano, J., Castro-Nuño, M., Pedregal, D., 2010. An econometric analysis of the effects of the penalty points system driver's license in Spain. Accid. Anal. Prev. 42 (4), 1310–1319.

Chetty, R., 2012. Bounds on elasticities with optimization frictions: a synthesis of micro and macro evidence on labor supply. Econometrica 80, 969–1018.

Chetty, R., Friedman, J.N., Olsen, T., Pistaferri, L., 2011. Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: evidence from Danish tax records. Q. J. Econ. 126, 749–804.

Chetty, R., Friedman, J.N., Saez, E., 2013. Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. Am. Econ. Rev. 103 (7), 2683–2721.

Chetty, R., Looney, A., Kroft, K., 2009. Salience and taxation: theory and evidence. Am. Econ. Rev. 99 (4), 1145–1177.

Congdon, W.J., Mullainathan, S., Kling, J.R., 2011. Policy and Choice: Public Finance through the Lens of Behavioral Economics. Brookings Institution Press, Washington, DC.

D'Haultfœuille, X., Durrmeyer, I., Février, P., 2016. Disentangling sources of vehicle emissions reduction in France: 2003–2008. Int. J. Ind. Organ. 47, 186–229.

DeAngelo, G.J., Hansen, B., 2014. Life and death in the fast lane: police enforcement and traffic fatalities. Am. Econ. J. Econ. Pol. 6 (2), 231–257.

De Bartolome, C., 1995. Which tax rate do people use: average or marginal? J. Public Econ. 56 (1), 79–96.

De Paola, M., Scoppa, M., Falcone, M., 2013. The deterrent effects of penalty point system in driving licenses: a regression discontinuity approach. Empir. Econ. 45 (2), 965–985.

Dusek, L., Traxler, C., 2017. Experience with Punishment and Specific Deterrence: Evidence from Speeding Tickets, Mimeo.

Ericson, K.M., 2011. Forgetting we forget: overconfidence and memory. J. Eur. Econ. Assoc. 9 (1), 43–60.

Feldman, N., Katuscak, P., Kawano, L., 2016). Taxpayer confusion: evidence from the child tax. Am. Econ. Rev. 106 (3), 807–835.

Finkelstein, A., 2009. E-ztax: Tax salience and tax rates. Q. J. Econ. 124 (3), 969–1010.

Gillitzer, C., Kleven, H., Slemrod, J., 2017. A characteristics approach to optimal taxation: line drawing and tax-driven product innovation. Scand. J. Econ. 119 (2), 240–267.

Goncalves, F., Mello, S., 2017. A few bad apples? Racial bias in policing. Princeton In: IRS Working Paper No. 608.

Graves, P., Lee, D., Sexton, R., 1993. Speed variance, enforcement, and the optimal speed limit. Econ. Lett. 42, 237–243.

Hansen, B., 2015. Punishment and deterrence: evidence from drunk driving. Am. Econ. Rev. 105 (4), 1581–1617.

Jondrow, J., Bowes, M., Levy, R., 1983. The optimal speed limit. Econ. Inq. 21, 325–336.

Kleven, H., 2016. Bunching. Annu. Rev. Econ. 8, 435–464.

Kleven, H., Waseem, M., 2013. Using notches to uncover optimization frictions and structural elasticities: theory and evidence from Pakistan. Q. J. Econ. 128 (2), 669–723.

Lave, C., 1985. Speeding coordination, and the 55 MPH limit. Am. Econ. Rev. 75, 1159–1164.

Liebman, J., Zeckhauser, R., 2004. Schmeduling. In: Working Paper. Harvard University.

Montag, J., 2014. A radical change in traffic law: effects on fatalities in the Czech Republic. J. Public Health 36, 539–545.

Peden, M., Scureld, R., Sleet, D., Mohan, D., Hyder, A., Jarawan, E., Mathers, C., 2004. World Report on Road Traffic Injury Prevention. World Health Organization, Geneva.

Rasmusen, E., 1995. How optimal penalties change with the amount of harm. Int. Rev. Law Econ. 15, 101–108.

Saez, E., 2010. Do taxpayers bunch at kink points? Am. Econ. J. Econ. Pol. 2, 180–212.

Sallee, J., Slemrod, J., 2012. Car notches: strategic automaker responses to fuel economy policy. J. Public Econ. 96, 981–999.

Slemrod, J., 2013. Buenas notches: lines and notches in tax system design. eJ. Tax Res. 11, 259–283.

Taubinsky, D., Rees-Jones, A., 2017. Attention variation and welfare: theory and evidence from a tax salience experiment. Rev. Econ. Stud. http://dx.doi.org/10.1093/restud/rdx069.

van Bentham, A., 2015. What is the optimal speed limit on freeways? J. Public Econ. 124, 44–62.