I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13639

Optimal Model Selection in RDD and Related Settings Using Placebo Zones

Nathan Kettlewell
Peter Siminski

AUGUST 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13639

Optimal Model Selection in RDD and Related Settings Using Placebo Zones

Nathan Kettlewell
University of Technology Sydney and IZA

Peter Siminski
University of Technology Sydney and IZA

AUGUST 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Optimal Model Selection in RDD and Related Settings Using Placebo Zones*

We propose a new model-selection algorithm for Regression Discontinuity Design, Regression Kink Design, and related IV estimators. Candidate models are assessed within a 'placebo zone' of the running variable, where the true effects are known to be zero. The approach yields an optimal combination of bandwidth, polynomial, and any other choice parameters. It can also inform choices between classes of models (e.g. RDD versus cohort-IV) and any other choices, such as covariates, kernel, or other weights. We use the approach to evaluate changes in Minimum Supervised Driving Hours in the Australian state of New South Wales. We also re-evaluate evidence on the effects of Head Start and Minimum Legal Drinking Age. We conclude with practical advice for researchers, including implications of treatment effect heterogeneity.

JEL Classification: C13, C52, I18

Keywords: regression discontinuity, regression kink, graduated driver licensing

Corresponding author:
Nathan Kettlewell
Economics Discipline Group
University of Technology Sydney
Sydney, NSW
Australia
E-mail: Nathan.Kettlewell@uts.edu.au

* For useful discussions and comments on earlier drafts, we especially thank Marc Chan, Mengheng Li and Timothy J Moore, as well as Victoria Baranov, Colin Cameron, Yingying Dong, Denzil Fiebig, Mario Fiorini, Christopher Taber, and participants at seminars at University of Technology Sydney, University of Sydney, University of New South Wales and the Melbourne Institute for helpful feedback and discussions. Any errors are our own.

1 Introduction

Policy rules frequently create discontinuous 'jumps' in exposure to policies and programs. The regression discontinuity design (RDD) has become a key tool for empirical researchers in these settings (see e.g. Imbens & Lemieux, 2008; Lee & Lemieux, 2010; Cattaneo et al., 2020, for overviews). In the canonical sharp RDD case, the treatment T changes discontinuously from $T = 0$ to $T = 1$ at some threshold along the running variable X. Setting $x = 0$ as that threshold, the goal is to estimate the change in the outcome Y at $x = 0$:

$$\tau(x) = \lim_{x \to 0^+} E[Y|X = x] - \lim_{x \to 0^-} E[Y|X = x] \tag{1}$$

$\tau(x)$ is commonly estimated by local polynomial regression. Researchers select some neighbourhood of observations around $x = 0$ (the bandwidth) where $E[Y_{T=0}|X = x]$ and $E[Y_{T=1}|X = x]$ are expected to meet the continuity assumption (Hahn et al., 2001) and estimate the jump in Y while flexibly controlling for X above and below $x = 0$.

RDD is appealing because it facilitates estimation of causal effects under relatively weak assumptions. Moreover, the assumptions for RDD have simple, testable implications (see e.g. McCrary, 2008; Cattaneo et al., 2019). The ability to visualize RDDs in simple plots of the running variable and outcome (Calonico et al., 2015) also gives an appealing air of transparency to this approach. A number of related estimators extend the basic RDD. The regression kink design (RKD) identifies causal effects by exploiting discontinuous changes in the slope of the running variable under similar assumptions to RDD (Card et al., 2015). Fuzzy RDD and RKD deal with situations where only the probability of treatment changes at the threshold, or treatment is a continuous variable. Dong (2018) suggests a regression probability jump and kink design (RPJKD) for settings where there is both a discontinuous 'jump' and/or 'kink'. In many related settings it is also possible to fit a global polynomial through the running variable and instrument the treatment using binned means (cohort-IV),

as in Angrist and Lavy (1999).[1]

While theoretically appealing, when using discontinuity designs researchers face a daunting challenge in selecting a preferred estimator. The choice of bandwidth involves a difficult trade-off between bias and variance. Researchers must also choose what order of polynomial to use, what kernel to use, whether to include covariates, and in some applications what discontinuity model to estimate (e.g. in situations where there is a 'jump' and a 'kink'). Sometimes it is also useful to adopt different polynomial orders on the left and the right of the threshold, or different bandwidths. Consequently, researchers will typically have thousands of potential estimators to select from, and there is no widely accepted standard for making this choice. In a given application, estimates may vary widely depending on the choices the researcher makes.

Various solutions to model selection have been suggested, but these typically focus on one decision and fix other important decisions. Optimal bandwidth selection has received a lot of attention. In economics, least squares cross validation and plug-in approaches have dominated (Imbens & Lemieux, 2008). Cross validation methods typically select a bandwidth to minimizes the mean squared error of the local polynomial fit. Ludwig and Miller (2005, 2007) discuss an alternative approach that minimizes error at the boundary, although ultimately reject this method for their application. Early plug-in approaches also focused on the polynomial function's fit, for example the rule-of-thumb approach discussed in Fan and Gijbels (1996) (see also Lee & Lemieux, 2010). In an influential paper, Imbens and Kalyanaraman (2012) (IK) argue that instead of focusing on the global fit of the polynomial function, the bandwidth should minimize the asymptotic mean squared error of the treatment effect (boundary) estimator (see also Ludwig & Miller, 2005). They derive a plug-in algorithm that selects the optimal bandwidth to achieve this. Calonico et al. (2014) (CCT) derive a bias correction to IK's method to improve confidence interval estimation. The

[1]The cohort-IV approach can also be used in related situations where there is no clear discontinuity, and yet treatment is a non-smooth function of the running variable. See for example Imbens and van der Klaauw (1995), Bound and Turner (2002) and Cousley et al. (2017).

IK/CCT approach is now popular in applied work.

Card et al. (2017), however, caution against using the IK/CCT approach as a default and demonstrate through simulations that it may not perform best for a given application.[2] Further, none of these approaches deals with the simultaneous modelling choices researchers need to make. For example, the optimal bandwidth will almost always depend on the polynomial order.[3] There has been less theoretical development on the question of polynomial choice. Gelman and Imbens (2019) argue that researchers should generally use local linear or quadratic regressions because higher order terms can induce undesirable effects on the estimates.[4] Pei et al. (2020) are less critical of higher order terms and suggest that, conditional on a given bandwidth and other modelling choices, researchers should calculate the implied asymptotic mean squared error for the boundary estimator (similar in spirit to IK/CCT for bandwidth selection). They also show, through a review of recent literature, that most researchers simply default to using local linear estimation.

To understand how researchers are dealing with the challenges of model selection in discontinuity designs, we conducted a review of papers published in leading journals for applied economics research in 2019 (See Table 1 and Appendix Table D1).[5] Of the 26 papers we identified, 12 gave no formal rationale for their preferred bandwidth. Of those that did motivate their choice, 13 used IK/CCT; however, many of these merely used the method to 'guide' their choice (e.g. by noting that the IK/CCT bandwidth was similar to whatever bandwidth they ultimately used). Almost all studies conducted some kind of sensitivity

[2]Card et al. (2017) suggest that the regularization term used in the IK/CCT plug-in approach may be overly punitive to large bandwidths in practice. The regularization term is used to account for the fact that the curvature parameters for the polynomial fits – which are parameters themselves in the plug-in formula – are unknown and must be estimated from the data.

[3]Hall and Racine (2015) propose a leave-one-out cross validation approach that jointly selects the bandwidth and polynomial order. Their cross validation approach is subject to the issues discussed in IK.

[4]Gelman and Imbens (2019) point out that higher order terms can have the practical effect of giving disproportional weighting to certain observations, are generally not selected on the basis of optimizing the objective of boundary estimation, and can lead to misleading inference.

[5]We searched Econlit on 30 April 2020 using the terms "discontinuit*", "fuzzy RD" and "regression kink" (contained anywhere) and restricted results to the following journals: *AEJ: Applied Economics*; *AEJ: Economic Policy*; *American Economic Review*; *Journal of Health Economics*; *Journal of Human Resources*; *Journal of Labour Economics*; *Journal of Political Economy*; *Journal of Public Economics*; *Quarterly Journal of Economics*; *Review of Economic Studies*; and *Review of Economics and Statistics*.

Table 1: Discontinuity studies published in leading journals in 2019

	Sharp RDD	Fuzzy RDD	Cohort-IV
Papers using this model	15	10	2
Method for bandwidth choice			
No stated method	6	4	2
IK/CCT	8	6	0
Method for polynomial choice			
No stated method	10	9	2
Local linear polynomial as baseline	11	8	-
Robustness tests			
Varied bandwidth	15	9	1
Varied polynomial	13	5	1

Notes: One paper used both sharp and fuzzy RDD as main specifications, so columns add to more than the sample size ($n = 26$). Papers that use spatial or multivariate RDD are included in Sharp RDD or Fuzzy RDD (depending on whether the treatment had complete or partial take-up). No papers used RKD; however, one cohort-IV study did use kink variation as an instrument. See Appendix Table D1 for a more detailed overview.

testing by varying the bandwidth. Only five studies provided any justification for their chosen order of polynomial. Most studies (18) used local linear regression and typically added higher order terms as a robustness check.[6]

Overall, we surmise that there is no consensus among applied researchers about how to select a preferred model in discontinuity settings. In many cases, researchers seem to be selecting a baseline model either arbitrarily or based on possible defaults like local linear regression. The focus away from emphasizing a preferred model and towards sensitivity analysis may be problematic in certain applications. For example, if the true data generating process (DGP) for the running variable is quadratic, then estimates may be sensitive to local linear estimation. But the reverse will not be true (higher order terms will simply decrease precision but not bias estimates if the DGP is linear). This could lead to discounting of evidence from studies with non-linear DGPs. Further, if there is no clear preferred model, there is also no clarity around the confidence interval for the treatment effect. More generally, the emphasis on robustness tests means that we may be 'setting the bar too high' for what constitutes credible evidence in RDD and related contexts.

[6]Other common robustness checks included adding covariates, different kernels, 'donuts' around the threshold and falsification tests using placebo cut-off points.

In this paper we propose a new method for model selection with broad application. Our method allows researchers to select an optimal combination of bandwidth, polynomial, and any other choice parameters he/she wants to consider. It can also be used to choose between competing models (e.g. RDD versus cohort-IV) in certain settings and can accommodate non-linear dependent variables. It relies on using observations of the running variable away from the discontinuity (the placebo zone) as a training ground to assess the performance of candidate models where a 'pseudo-treatment' effect is known to be zero. The estimator that minimizes the preferred performance criterion (e.g. lowest root mean squared error) across all pseudo-treatments is then selected as the 'best' specification for estimating the actual treatment effect. Our approach is applicable in settings where the point of discontinuity can be reasonably thought of as being randomly chosen from the domain of the running variable..

We are not the first to recognize the value in placebo zone data. Imbens and Lemieux (2008) suggest testing for jumps at specific psuedo-thresholds as a general test for specification error, a common practice in applied work. Wing and Cook (2013) use the placebo zone to create a kind of differences-in-differences structure, which they argue can improve precision and allow one to learn something about the treatment effect away from the threshold. Gelman and Imbens (2019) use results from the distributions of placebo estimates to inform general advice about higher order terms in RDD studies. Closest to our work is Ganong and Jäger (2018), who suggest a randomization inference approach to hypothesis testing based on the distribution of pseudo-treatment effect estimates (we propose extensions to this procedure). We extend all of this work by using the placebo zone for *ex ante* model selection, which to the best of our knowledge is a new idea.

Our approach also has parallels with studies that use estimates from randomized controlled trials (RCTs) to assess RDD estimators. A prominent example is Hyytinen et al. (2018) who use one such RCT-RDD pair, and conclude that CCT bias-corrected estimators perform well in that application.[7] In this literature, as in our approach, the assessment

[7]See Chaplin et al. (2018) for a review and meta-analysis of similar studies.

rests on knowing the true parameter that the RDD estimator targets. In that literature, the target is the estimate generated by an RCT. In our case, the target estimate is zero, since there is no actual treatment. Our approach builds on the RCT-RDD approach in three important ways. First, rather than making a single comparison of RDD to RCT estimates, our approach assesses each candidate estimator's performance repeatedly – at hundreds or thousands of placebo thresholds throughout the placebo zone. Second, these comparisons serve to inform the choice of estimator to apply within the same context, to estimate the effect of a real treatment using the same data, in a range of the running variable that borders the placebo zone. There is no reason to believe that the best-performing estimator will perform best in other unrelated contexts where the DGP may be completely different, or with other sources of data. Thirdly, the target parameter in the RCT-RDD literature is subject to sampling bias, whereas the placebo-zone target of zero is known with certainty.

We demonstrate our approach with a novel evaluation of a policy designed to reduce motor vehicle accidents (MVAs) for young drivers. The policy requires that learner drivers meet a minimum supervised driving hours (MSDH) mandate before being able to drive independently; a common requirement in jurisdictions using graduated driver licencing systems.[8] We are among the first to causally evaluate the effect of MSDH on MVAs.

In New South Wales, Australia, policy rules created two discontinuities whereby young drivers needed to complete either 0, 50 or 120 MSDH depending on their birth cohort and date of obtaining license. This setting is particularly interesting for demonstrating our method because we can use it to not only select model parameters, but also *model type*. There are apparent first-stage discontinuities in both the level and slope of treatment. We could estimate a global polynomial model like Angrist and Lavy (1999) (cohort-IV), RKD, RDD, or RPJKD, and it is *a priori* unclear which approach we should adopt. Further, within each of these models we need to make important functional form and bandwidth choices. In this setting, there are also good reasons to consider models with both asymmetric bandwidths

[8]Countries implementing strict graduated driver licensing systems include Australia, Canada, New Zealand and the U.S. See L. J. Bates et al. (2014) for a broad overview of these systems.

and asymmetric polynomial orders. Institutional details prevent long bandwidths on the left (but not on the right) of the threshold. Institutional details also result in complete compliance on the right, but strong non-linearity on the left, of the threshold. In total we consider almost 10,000 different estimators considering model type, functional form and bandwidth.

Somewhat surprisingly, our 'best' estimator is a month-of-birth cohort-IV with linear trend. A mixed order RPJKD also performs well, and indeed performs best when asymmetric bandwidths are allowed. Strikingly, the root mean squared error is about five times greater across the placebo zone if we use the bandwidths suggested by CCT rather than our preferred bandwidths. In a different application, Card et al. (2017) come to a similar conclusion, drawing on Monte Carlo simulations.

We find that going from 0 to 50 MSDH lowers the probability of an MVA in the first year of independent driving by 1.4 percentage points (21%). This estimate is robust to a randomization inference procedure similar to Ganong and Jäger (2018), even after adjusting for serial correlation in the distribution of the placebo estimates. In further analysis we find that the reduction in MVAs is not driven by people delaying their licensing, is similar magnitude if we restrict attention of more serious MVAs, is experienced by both males and females, and disappears in the second year of independent driving. We also find that going from 50 to 120 MSDH does not lower MVAs, which is consistent with strongly diminishing returns at this level.

To further demonstrate our approach, we re-evaluate evidence on the effect of Head Start on child mortality (Ludwig & Miller, 2007) and the minimum legal drinking age on drinking behavior (Lindo et al., 2016). For both applications, the best performing model is linear RDD with a relatively long bandwidth (much longer than in the original papers) and CCT estimators perform considerably worse than our best models in the placebo zone.

We recommend that researchers consider using our approach whenever feasible. This means settings where the researcher has access to a sufficiently wide placebo zone and where

the threshold is plausibly random with respect to the underlying DGP (we provide guidance on how this assumption could be assessed in practice). We think these conditions would be met in a great number of discontinuity settings; in fact, settings where there are insufficient placebo observations can be thought of as the subset of discontinuity studies where data constraints rule out consideration of large bandwidths.

The remainder of the paper is structured as follows. Section 2 describes the details of our main application and Section 3 describes the data. In Section 4 we illustrate in detail how the placebo-zone approach is applied in our context. Section 5 presents results which adopt the chosen estimators. Section 6 presents a re-evaluation of Head Start and minimum legal drinking age studies. Section 7 concludes and discusses practical considerations and recommendations for using the placebo zone approach.

2 An application: Minimum supervised driving hours and motor vehicle accidents

2.1 Overview

Globally, MVAs are the leading cause of death for children and young adults, with more than 1.3 million people aged 5-29 years dying from MVAs each year (WHO, 2018). To reduce the fatality rate for young drivers, governments around the world have introduced graduated driver licensing (GDL). GDL limits the exposure of young drivers to risky situations with the goal of better preparing them for unsupervised driving. It typically operates in three stages: a learner stage in which driving is supervised; a provisional stage in which driving is unsupervised but subject to restrictions; and an unrestricted stage. To progress, drivers are required to demonstrate competence by passing written exams and practical driving tests.

During the learner stage drivers usually need to complete a mandatory number of supervised driving hours – the MSDH requirement. Most U.S. states mandate between 40-60

hours (IIFHS, 2020). In Australia, the three most populous states (New South Wales (NSW), Victoria and Queensland) require 100-120 hours.

It is generally believed that GDL as a system has reduced MVAs for young drivers (McKnight & Peck, 2002; Foss, 2007; Shope, 2007); however, there is little evidence on the independent effects of different components of GDL. Typically researchers rank GDL systems by some measure of 'strictness' and use state variation in regulatory settings to identify policy effects (e.g. Dee et al., 2005; Chen et al., 2006; Traynor, 2009; Trempel, 2009; Karaca-Mandic & Ridgeway, 2010; Masten et al., 2011; Lyon et al., 2012; Steadman et al., 2014).[9] Results consistently show that states with stricter GDL systems experience lower rates of fatalities and MVAs involving injury among teenage drivers.

We are aware of only one study (Gilpin, 2019) that attempts to estimate the independent *causal* effect of MSDH on MVAs.[10] Gilpin (2019) uses a difference-in-differences design with variation between and within U.S. states and finds going from no MSDH requirement to having some MSDH requirement counter-intuitively increased fatalities overall, but had no effect per licensee.

2.2 Policy environment and causal variation

NSW adopted GDL on 1 July 2000. Prior to this, a licensing system with GDL features operated. Under the pre-July 2000 system, the minimum age for obtaining a learner license was 16 years, there was a minimum six-month learner period and one year provisional license period, and the minimum age for obtaining a provisional license was 17 years. There was no MSDH requirement. The introduction of GDL resulted in two restricted provisional license periods – provisional 1 (P1) and provisional 2 (P2), which remain in place today. It also resulted in a large increase in MSDH – from 0 to 50 hours. Importantly, the six-

[9]Moore and Morris (2020) identify the causal effect of one common component of GDL – night-time passenger restriction – on MVAs in NSW, Australia. Using variation in MVAs by time-of-day and a difference-in-differences design, they find large reduction effects.

[10]Trempel (2009) and McCartt et al. (2010) estimate models that control for MSDH in U.S. state-level studies but do not control for state fixed-effects or time trends. O'Brien et al. (2013) study an increase from 0 to 30 MSDH in Minnesota using a before-after design.

month minimum learner period and 17 years minimum age for obtaining a provisional license remained in place.

Although the 1 July 2000 MSDH increase was not implemented in isolation, it was implemented in such a way that people born up to one year before 1 July 1984 experienced the same provisional licensing conditions as those born after this date. This is because people born within one year prior to 1 July 1984 turned 16 before the introduction of GDL (meaning they could obtain their learner license before 1 July 2000 and avoid the increase to MSDH) but turned 17 *after* 1 July 2000, meaning they could not avoid the new GDL provisional regulations. Consequently, the GDL experience of people born within one year of 1 July 1984 only differs with regards to the 50 MSDH requirement.

The GDL system was expanded on 1 July 2007. The most significant changes were passenger restrictions for night-time driving for P1 drivers, a zero-tolerance policy for speeding (immediate three-month suspension of license) and an increase in MSDH from 50 to 120 hours (minimum 20 hours night-time driving). There was also an increase to the minimum learner period from six to 12 months. In Table 2 we highlight the main difference between the pre- and post-July 2007 regimes (see L. Bates, 2012, for a detailed comparison). As with the 1 July 2000 policy changes, the 17 years minimum age for obtaining a provisional license meant that people born up to one year prior to 1 July 1991 (meaning they would turn 17 *after* 1 July 2007) could obtain their learner license before 1 July 2007 and avoid the MSDH increase but would be subject to the same provisional regulations as those born after 1 July 1991.[11]

Our empirical analysis exploits the fact that people born just before 1 July 1984 (1991) are likely to be statistically similar to those born just after 1 July 1984 (1991) but differ in their MSDH experience. Since people often delay getting their license until sometime after their 16th birthday, there is a positive slope in the probability of treatment on the

[11]Because the minimum learner period also increased at the same time as MSDH in 2007, these policy effects may be confounded in our analysis. To separate these effects and isolate the impact of increased driving practice, we consider the impact of the policy change on time spent on the learner license and how our results change when we control for this.

Table 2: NSW GDL characteristics

	1 July 2000—30 June 2007	From 1 July 2007
MSDH	50	120[a]
Min. learner age	16 years	16 years
Min. learner period	6 months	1 year
Min. P1 age	17 years	17 years
Min. P1 period	1 year	1 year
P1 restrictions	Max speed (90km/h); 4 demerit points[b]; engine restrictions[c]	Max speed (90km/h); 4 demerit points; engine restrictions[c]; night-time passenger restrictions; immediate license suspension for speeding
Blood alcohol limit	0.02 (0.00)[d]	0.00

Notes: [a]In NSW drivers receive financial penalties and demerit points for driving offences. Drivers who accrue a critical number of demerit points have their license suspended (4 for P1 drivers, 12 for unrestricted license drivers). [b]Minimum 20 hours at night. In December 2009 new rules were introduced that allowed learners to convert hours with a qualified driving instruction at a ratio 3:1 with regular supervised driving (limited to 10 hours). [c]Since 11 July 2005 P1 drivers have been prohibited from driving certain high-powered vehicles. [d]Lowered to this on 3 May 2004.

left-hand-side of the threshold, while everyone is treated on the right-hand-side.

2.3 Compliance

We do not observe learner driving hours so cannot assess compliance directly. However, the limited Australian evidence supports high compliance with MSDH regulations. For example, surveys of newly licensed drivers found 98.2% complied with NSW's 50 MDSH requirement (L. Bates et al., 2010), while only 12.8% admitted to rounding up hours and 4% to including additional hours not undertaken in Queensland (Scott-Parker et al., 2011). A survey by L. Bates et al. (2014) also found strong agreement from parent supervisors about the accuracy of recorded hours.

One reason to expect high compliance in our setting is because learner drivers are required to record all journeys in a log book, with each entry signed off by the supervising driver (not a P1 or P2 driver). If there is evidence of falsification the learner may be barred from taking the practical driving test for up to six weeks and fined (fines also apply to supervising

drivers).

A related question is whether the policy is binding at all. L. Bates et al. (2010) compared new drivers in NSW to Queensland when NSW had a 50 MSDH requirement and Queensland had none. The average self-reported hours was only slightly higher in NSW (73 compared to 64). However, while 98.8% of drivers reported completing at least 50 hours in NSW, more than half in Queensland reported doing less than this. A 50 MSDH requirement would have therefore been binding for a significant portion of learners in Queensland. It is important to note that any effects of increased MSDH we estimate will be driven by the subsample of learners who would have completed less than the minimum requirement in the absence of the policy.

3 Data

Our data are individual level administrative records supplied by the NSW Centre for Road Safety (CRS). Driver licencing data come from the universe of licensing history for NSW drivers born from 1 January 1980. For these individuals, we know their age in (completed) weeks at the time they obtain their license. The MVA data are from a separate dataset containing the universe of police reported MVAs from 1 January 1996 to 26 October 2017. MVAs are accidents occurring on NSW roads in which at least one vehicle was towed away or one of the occupants was injured or killed, which by law must be reported to NSW Police (we exclude motorcycle crashes from the analysis). We link the license and MVA datasets using a unique identifier provided by CRS. Our study received ethics approval from the UTS Human Research Ethics Committee (Application number ETH17-1547).

3.1 Main variables

Our outcome variables are indicators for whether an MVA occurred within certain periods. We focus primarily on the probability a person was the driver in an MVA within one year

of obtaining his/her P1 license (during which drivers are typically aged 17-20 years). The one year criterion matches the mandatory time period before a P1 driver can take the test to become a P2 driver, and therefore reflects an expected period of progression in driver safety. We also find little evidence that the MSDH reforms improve driver safety beyond this period. In further analysis we limit attention to MVAs that resulted in injury to a driver or passenger or resulted in fatality.

Our running variable, date of birth (DOB), is constructed as follows. For each entry a person has in the license dataset (for example, when they renew their license or move to a different license class), we observe that person's age in weeks on that day. That means that for people with one entry, we know their precise DOB within 6 days. For people with multiple entries we can narrow that window down; for more than 50% of people we can narrow it down to within three days. We use the midpoint of the minimum and maximum possible DOB as our variable, considering all licensing history data available to us.

See Appendix Figure A1 for density plots for DOB. Further details on how the data are constructed are provided in Appendix B.

3.2 Descriptive Statistics

Sample means for the main variables in our study are in Table 3. We focus on two birth cohorts, centred ±365 days from the key dates for our two policy reforms. In Appendix Figure A2 we plot the variables by DOB for all years.

MVA incidences are generally lower for younger birth cohorts, although this is weaker for more serious MVAs that involve injury. For the circa 1 July 1984 cohort, the probability of any MVA within 12 months of obtaining P1 is 5.7%. This falls to 3.8% for the next 12 months, consistent with young drivers becoming safer with age and experience. The average age at which people obtain their learner license is 17 years, a full 12 months later than they become eligible. However, the mass of observations are just after the 16th birthday (Appendix Figure A3). Most people obtain their license shortly after they become eligible.

Table 3: Sample means by birth cohort

	1 July 1983–30 June 1985	1 July 1990–30 June 1992
MVA 1-year	0.057	0.044
MVA 1-2 years	0.038	0.028
Injury 1-year	0.022	0.020
Fatality 1-year	<0.000	<0.000
Age got L's	16.970	16.700
Age got P1	18.451	18.538
n	154,524	160,301

Notes: this table shows sample means of the key variables for observations within each of the two 'treatment zones' –. i.e. people born within one year of 1 July 1984, and 1 July 1991, respectively

Similarly, there is a large mass who obtain their P1 license shortly after their 17th birthday, while the average is 18.5 years for both birth cohorts.

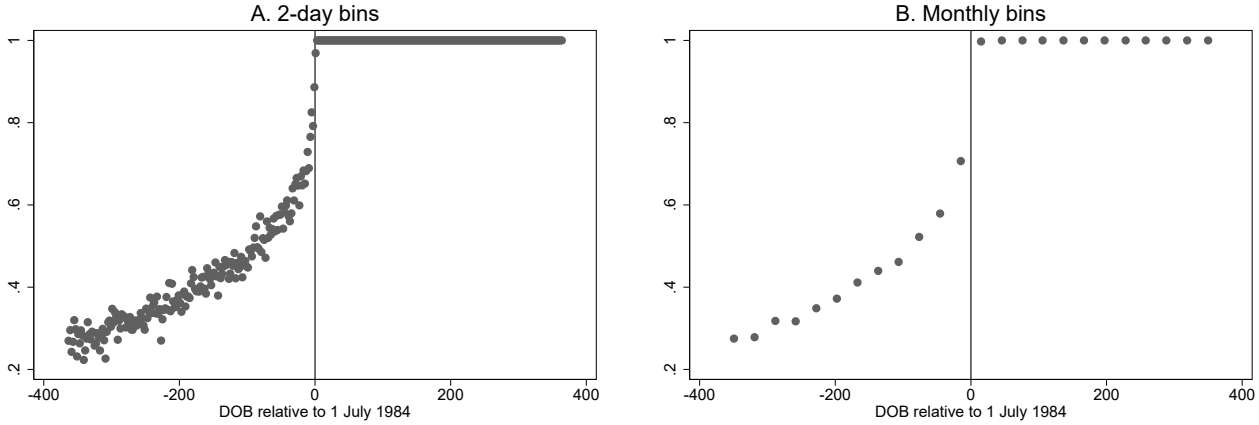
4 Model selection using a 'placebo zone'

The main aim of this paper is to demonstrate a new approach for model selection. In this section, we describe this approach in detail. We do this in the context of our application – estimating the effectiveness of learners' permit policy changes in NSW. We begin by describing the many credible candidate models which could be applied to estimate the effect of the policy changes. We then describe the 'placebo zone'. This is a set of 2,556 consecutive DOBs (from 1 July 1984 to 30 June 1991). Within this zone, there is no reason to suspect any systematic relationship between DOB and the outcome variables (MVAs). It therefore provides an opportunity for testing the performance of candidate models in estimating the true treatment effect within this zone (which is zero). Next, we summarize the performance of the candidate models within this zone. We then describe how the placebo zone estimates can also be used for an alternate inference procedure. Finally, we also consider the implications of various types of treatment effect heterogeneity which we impose into the placebo zone data.

4.1 Candidate models

The first-stage relationship between DOB and holding a 'new' learner's permit following the 2000 reform (0 to 50 MSDH) is shown in Figure 1 (see Appendix Figure A4 for the 2007 reform, and Figure A5 for scatter plots showing the reduced form for both reforms). Both panels draw on the same underlying data, differing only in the bin-size used in the plots. Panel A uses a 'small' bin-size of 2 days, while Panel B uses a 'large' bin size of 30 days.

Figure 1: First-stage relationship between DOB and 50 MSDH treatment



Notes: Both panels shows the mean value of the 'treatment' variable, by DOB. Treatment is defined as obtaining a first learner's permit on or after 1 July 2000, thereby subject to different MSDH requirements. The only difference between panels is the size of the DOB 'bins'.

Figure 1 shows complete compliance to the right of 1 July 1984.[12] It was not possible for anyone born on this date or after to hold an 'old' learner's permit due to administrative rules. The pattern on the left side is more complicated. Both panels show a monotonic upward, non-linear pattern. Panel B suggests the presence of a discontinuity at the threshold. In contrast, Panel A suggests no discontinuity, but a kink, caused by a very steep rise on the left side of the threshold.

This figure illustrates that many different estimators could potentially be used to estimate the effect of the reform. Candidate estimators could exploit the apparent kink, or

[12]While there appears to be very minor non-compliance, this is due to imprecision around DOB, as described in Section 3. We drop observations where we are uncertain about treatment status in our regression analysis.

the approximate discontinuity, or both. Or, they could instead employ a between-cohort-IV strategy. Each approach could be implemented using various alternate functional form assumptions (i.e. orders of polynomial, which need not be the same on each side of the threshold). Finally, one can choose between many bandwidths.

We first consider a total of 4,634 alternate candidate specifications, each with symmetrical bandwidth around the threshold. This consists of 14 different models, estimated using each possible bandwidth in the range of 35 to 365 days. In principle, we could consider larger bandwidths as well. This is prevented by practical considerations in our application. People born before 1 July 1983 were eligible for driver's licenses which differed in other important ways. Therefore we need an estimator which does not use data on people born before that date, hence making 365 days the largest feasible bandwidth.

Denoting outcome (i.e. MVA 1-year indicator) for person i by Y_i, DOB by X_i (centred at zero around 1 July 1984), treatment (obtained learner's permit after policy change) by T_i and an indicator for DOB $\geq$ 1 July 1984 (1991) by D_i, the first 11 candidate models are fuzzy RDD, RPJKD and RKD estimators. Each of these can be treated as instrumental variable models, with the structural equation given by Eq. 2 and first-stage given by Eq. 3. Full details on the estimation equations are in Table 4.[13]

$$Y_i = \alpha + \beta T_i + f(X_i, D_i) + e_i \tag{2}$$

$$T_i = \pi_0 + f(X_i, D_i) + g(X_i, D_i) + \epsilon_i \tag{3}$$

1. Model 1 is a conventional (fully-interacted) linear RDD.

2. Model 2 is an RDD model with a linear fit on the right side of the threshold, and a quadratic on the left. This is motivated by the first-stage relationship in Figure 1,

[13]We only consider a uniform kernel in our application, although it would be straightforward to vary the kernel along with other modelling dimensions. In practice, the choice of kernel typically has little influence on the estimates (Lee & Lemieux, 2010).

characterized by a clearly nonlinear relationship on the left, and perfect linearity on the right. We refer to this as a 'mixed polynomial' specification.

3. Model 3 is a conventional (fully-interacted) quadratic RDD.

The next four candidate models exploit both the discontinuity and the kink for identification. These are RPJKD estimators of the following form:

4. Model 4 is a conventional (fully-interacted) linear RPJKD.

5. Model 5 is a quadratic RPJKD, in which the quadratic term is not interacted with the threshold indicator.

6. Model 6 is an RPJKD model with a linear fit on the right side of the threshold, and a quadratic on the left.

7. Model 7 is a fully-interacted quadratic RPJKD.

Four more candidate models adopt conventional Regression Kink Designs:

8. Model 8 is a conventional (fully-interacted) linear RKD.

9. Model 9 is a quadratic RKD, in which the quadratic term is not interacted with the threshold indicator.

10. Model 10 is an RKD model with a linear fit on the right side of the threshold, and a quadratic on the left.

11. Model 11 is a fully-interacted quadratic RKD.

The remaining three candidate models are month-of-birth cohort-IV models, which exploit between-cohort variation in the probability of 'treatment'. Denoting month-of-birth fixed effects by θ_m, for these models, the first-stage becomes:

$$T_i = \pi_0 + f(X_i) + g(X_i, \theta_m) + \epsilon_i \tag{4}$$

12. Model 12 assumes a linear secular relationship between DOB and the outcome variable.

13. Model 13 assumes a quadratic secular relationship between DOB and the outcome variable.

14. Model 14 assumes a cubic secular relationship between DOB and the outcome variable.

4.2 The placebo zone

The 'placebo zone' is the set of DOBs between 1 July 1984 and 30 June 1991, inclusive. There were no apparent major licencing policy changes which were likely to have affected MVAs in a way that depends on DOB within this zone.[14] Figure 2 shows the MVA rate by month of birth within this zone (in 30 day bins), with a lowess fit. Generally, the pattern is relatively smooth, with a slight downward trend, apart from perhaps the first 5 months.

Within this zone, we create placebo treatments in a way that mimics the true treatment selection process. For example, in the first placebo, persons are deemed treated if they obtained their license on or after 1 July 2001. The first-stage relationship between DOB and this placebo treatment is shown (in 2-day bins) in Figure 3, with a 365 day bandwidth around the DOB threshold of 1 July 1985. This relationship closely resembles the true treatment profile around the 1 July 1984 DOB, which we show in Figure 1. Similar patterns are found for the other placebo DOB thresholds in this zone.

After collapsing to DOB-level (and weighting by cell-size), we estimate the placebo treatment effect (which we know to be zero and constant across entities) using each of the 4,634 candidate models.[15] We repeat this for all 1,826 placebo treatment thresholds, and summarize the performance of each candidate model.

[14]Policy changes that may have affected MVAs in our window are lowering the Blood Alcohol Limit from 0.02 to 0 (3 May 2004), engine restrictions (11 July 2005) and changes to the GDL system that occurred on 1 July 2007 for P1 drivers (see Table 2), in particular night-time passenger restrictions. However, days exposed to these policies is a smooth function of DOB (see Appendix Figure A6).

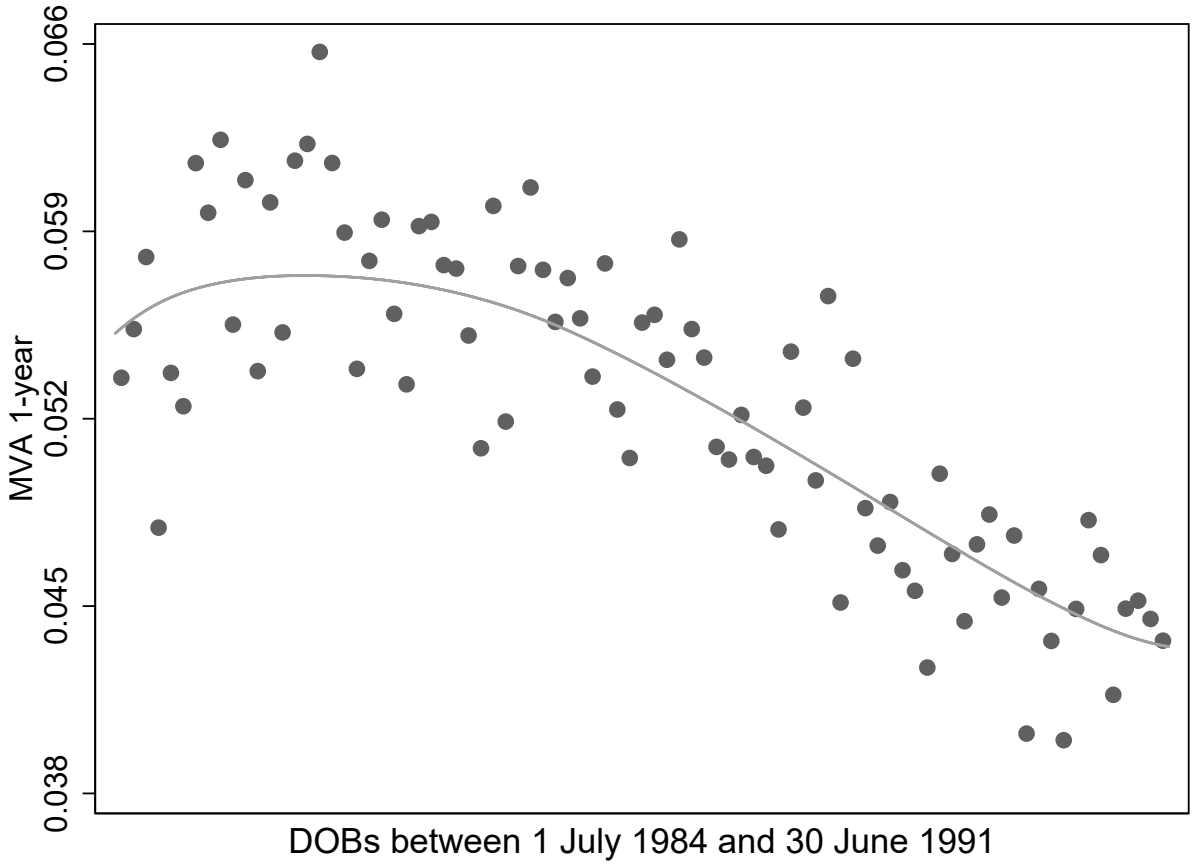
[15]The results are almost identical when uncollapsed microdata are used instead, but estimation is much faster with collapsed data.

Table 4: Candidate model equations

Model	Description	$f(.)$	$g(.)$
		RDD models	
1	RDD - linear	$f(.) = \gamma_1 X_i + \gamma_2 X_i D_i$	$g(.) = \pi_1 D_i$
2	RDD - mixed polynomial	$f(.) = \gamma_1 X_i + \gamma_2 X_i D_i + \gamma_3 X_i^2 (1 - D_i)$	$g(.) = \pi_1 D_i$
3	RDD - quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i D_i + \gamma_3 X_i^2 (1 - D_i) + \gamma_4 X_i^2 D_i$	$g(.) = \pi_1 D_i$
		RPJKD models	
4	RPJKD - linear	$f(.) = \gamma_1 X_i$	$g(.) = \pi_1 D_i + \pi_2 X_i D_i$
5	RPJKD - quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2$	$g(.) = \pi_1 D_i + \pi_2 X_i D_i$
6	RPJKD - mixed polynomial	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2 (1 - D_i)$	$g(.) = \pi_1 D_i + \pi_2 X_i D_i$
7	RPJKD - interacted quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2 (1 - D_i) + \gamma_3 X_i^2$	$g(.) = \pi_1 D_i + \pi_2 X_i D_i$
		RKD models	
8	RKD - linear	$f(.) = \gamma_1 X_i$	$g(.) = \pi_1 X_i D_i$
9	RKD - quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2$	$g(.) = \pi_1 X_i D_i$
10	RKD - mixed polynomial	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2 (1 - D_i)$	$g(.) = \pi_1 X_i D_i$
11	RKD - interacted quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^2 (1 - D_i)$	$g(.) = \pi_1 X_i D_i$
		Birth cohort-IV models	
12	Birth cohort-IV - linear	$f(.) = \gamma_1 X_i$	$g(.) = \theta_m$
13	Birth cohort-IV - quadratic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2$	$g(.) = \theta_m$
14	Birth cohort-IV - cubic	$f(.) = \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3$	$g(.) = \theta_m$

Notes: This table summarizes the functional forms of the models included in the placebo zone trials in our main application. The functions $f(.)$ and $g(.)$ are components of the full specifications shown in equations 2, 3, and for models 12-14, equation 4.

Figure 2: Trend in MVAs by DOB in the placebo zone



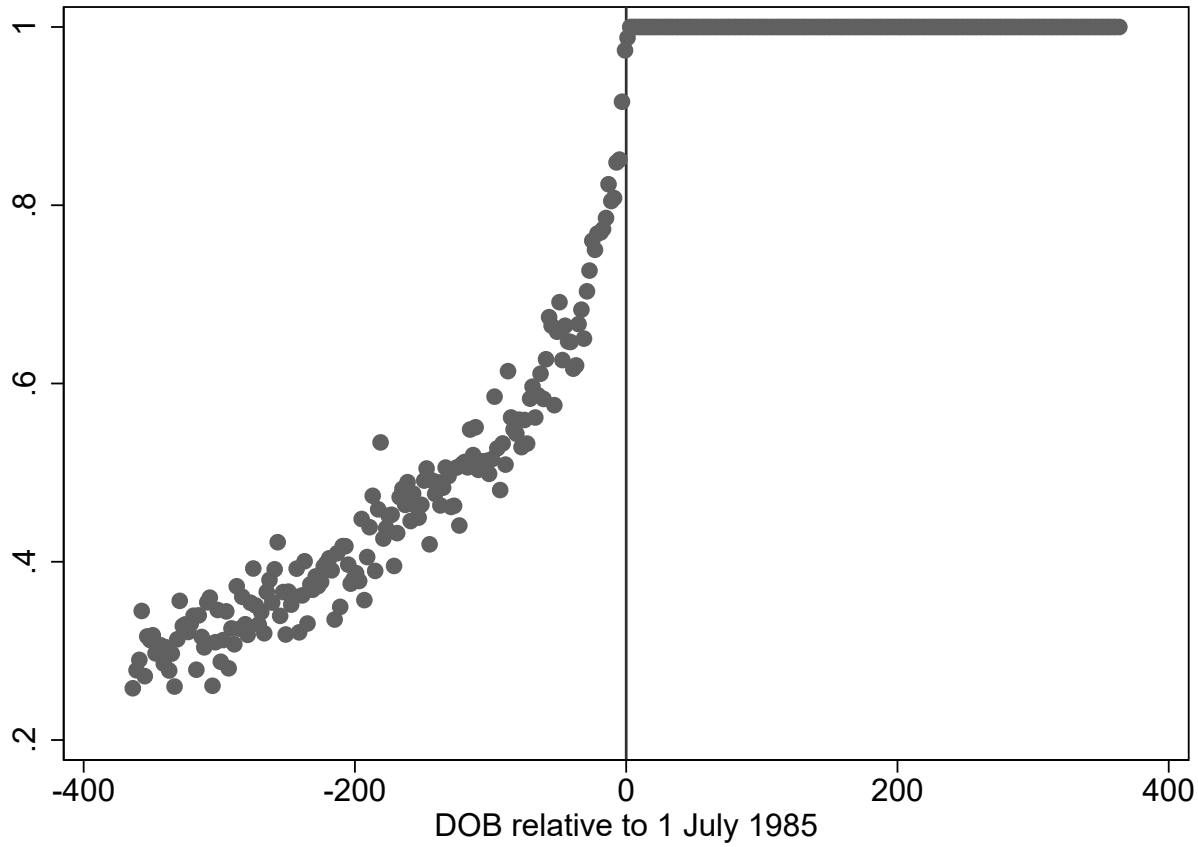
Notes: The plot shows the proportion of people who crashed within one year of receiving a provisional drivers license (the main outcome variable in the analysis) by DOB within the placebo zone. The plot uses 30-day bins of DOB.

4.3 Model performance in the placebo zone

Table 5 summarizes the performance of each candidate model. It would not be practical to report on the performance of all 4,634 candidates. Instead we show only the results for the bandwidth which yields the lowest root mean squared error (RMSE) for each model type. The first clear feature of this table is that for every model considered, large bandwidths (365 days in all but one case) yield the smallest RMSEs, compared with smaller bandwidths. Secondly, most models have appropriate coverage rates.

In our application, four models stand out with the lowest RMSE. The best performing model (RMSE = 0.0051) is 'Model 12' – the month-of-birth cohort-IV model with a linear

Figure 3: Placebo first-stage between DOB and obtained learner permit after 1 July 2001



Notes: This scatter plot is based on Figure 1 Panel A. Here, however, the range of DOB is shifted by one year, and so is the definition of 'treatment', which is a function of date received first Ls.

trend. This is closely followed by 'Model 6' (RMSE = 0.0052) – the RPJKD with mixed polynomial fit (quadratic on the left and linear on the right). Next are the RKD with mixed-polynomials (RMSE = 0.0057) and the linear RPJKD (RMSE = 0.0060). All four have similarly good coverage (at least 93.6%), and small average bias (0.001 at most).

We also consider a model-averaging approach. It is defined as the weighted average of the estimates from the 14 candidate models (each with full 365 day bandwidth). The weights are set to the inverse of the MSE of each candidate model. The performance of this weighted average estimator is also shown in Table 5. Whilst its performance is good, its RMSE is higher than Models 12 and 6. The coverage of this estimator is not shown as its variance has not been derived.

Table 5: Candidate model performance in the placebo zone

Model	Description	RMSE	Optimal BW	Coverage	Bias
1	RDD - linear	0.0083	365	0.962	-0.0004
2	RDD - mixed polynomial	0.0199	365	0.921	0.0014
3	RDD - quadratic	0.0230	365	0.927	0.0015
4	**RPJKD - linear**	0.0060	365	0.936	0.0010
5	RPJKD - quadratic	0.0073	365	0.980	-0.0005
6	**RPJKD - mixed polynomial**	0.0052	365	0.992	0.0000
7	RPJKD - interacted quadratic	0.0132	365	0.938	0.0005
8	RKD - linear	0.0096	355	0.910	0.0028
9	RKD – quadratic	0.0179	365	0.953	0.0019
10	**RKD - mixed polynomial**	0.0057	365	0.984	0.0002
11	RKD - interacted quadratic	0.0177	365	0.950	0.0019
12	**birth cohort-IV - linear**	0.0051	365	0.946	0.0006
13	birth cohort-IV - quadratic	0.0070	365	0.987	-0.0007
14	birth cohort-IV - cubic	0.0124	365	0.937	0.0001
WA	Inv-MSE weighted average	0.0055	365	n.d.	0.0003
C1	RDD conventional	0.0340	117	0.966	0.0012
C2	RDD bias corrected	0.0443	117/184	0.939	0.0015
C3	RKD conventional	0.0346	136	0.997	0.0014
C4	RKD bias corrected	0.0415	136/202	0.999	0.0019

Notes: This table summarizes the performance of each candidate model within the placebo zone. The key statistic is the RMSE of estimated treatment effects. There are 1,826 treatment effect estimates for every model, one for each placebo-zone threshold. The true treatment effect is known to be zero throughout the placebo zone, so zero is the target parameter for every estimator. With the exception of WA and C1-C4, every candidate model is trialled repeatedly with symmetric bandwidths ranging from 30 to 365 days. For each model, results from the bandwidth which yields the lowest RMSE are shown. In addition to the 14 main models, the model labelled WA is an estimator which (for each placebo-zone repetition) uses the inverse-MSE-weighted average of the estimates from the 14 main models, using each of those model's respective optimal bandwidth. Unlike the other models, those labelled C1-C4 use a CCT bandwidth selection procedure and default settings in Stata's -rdrobust- command.

The final four rows of Table 5 summarize the performance of four estimators proposed by CCT, and implemented using Stata's -rdrobust- command. These are conventional, and bias-corrected estimates using RDD and RKD, respectively.[16] The 'optimal' bandwidths for these estimators are determined within rdrobust, rather than the placebo zone procedure that we adopt for the other estimators.[17] As seen in the table, these bandwidths are much

[16]The results shown are for models estimated on collapsed microdata. As with the other estimators considered, the results with collapsed (DOB) data are very similar.

[17]More precisely, the CCT bandwidths shown in Table 5 are the average of bandwidths selected by rdrobust through the placebo zone.

smaller than the others. The key result, however, is that the performance of these estimators, as measured by RMSE, is worse than any of the other candidate models, and an order of magnitude worse than the best performing candidate models. This is consistent with the findings of Card et al. (2017)'s RKD Monte Carlo simulations.

4.4 Incorporating asymmetrical bandwidths

In every model tested on the placebo zone thus far, we have followed conventional practice and imposed the same bandwidth on the left and right sides of the threshold. Here we explore whether model performance can be improved by allowing for asymmetric bandwidths.

In particular, we have so far capped the bandwidth at 365 days on each of the thresholds. This is motivated by practical constraints in our application. Any more than 365 days to the left of the 1 July 1984 threshold would take us into territory where other important policy changes were implemented in a way that relates systematically with DOB. But we do not have the same issue on the right side of the threshold. Similarly, for the 1991 threshold, we have no constraints in the left side, though data constraints prevent us from considering bandwidths greater than 365 days on the right.[18]

We now repeat the placebo-zone model selection procedure for all 14 candidate models using two similar procedures.

1. We fix the bandwidth to 365 days on the left, whilst allowing the bandwidth to vary between 365 days and 730 days on the right. This will be informative for model selection in our analysis of the 2000 reform. The number of placebo thresholds in this exercise is 1,461, due to the need to include a larger maximum bandwidth.

2. We fix the bandwidth to 365 days on the right, whilst allowing the bandwidth to vary between 365 days and 730 days on the left. This will be informative for model selection

[18]The constraint is due to the fact that drivers who obtain their P1 license after age 25 are dropped from the sample because they are not required to meet the MSDH requirement (see Appendix B). We cannot impose this constraint consistently on the RHS of the 2007 reform because the end-date for our license data mean we do not always observe whether people got their P1 license by age 25.

in our analysis of the 2007 reform. The number of placebo thresholds is 1,461.

The results for Version 1 of this exercise are summarized in Table 6. It shows that performance is improved considerably for every model by allowing larger bandwidths on the right. In some cases, RMSE is reduced by more than 50%. The optimal right-side bandwidth varies considerably, from 550 up to the 730 day limit. Model 6 is the best performing model, with an optimal RHS bandwidth of 550 days. This is the best performing estimator amongst all candidates for estimating the effect of the 2000 reform. The weighted-average estimator, shown in the lowest row, does just as well as Model 6. Models 12, 4 and 10 continue to perform well.

Table 6: Candidate model performance in the placebo zone V1: Asymmetric bandwidths

Model	Description	RMSE	Optimal RHS BW	Coverage	Bias
1	RDD - linear	0.0069	550	0.958	-0.0005
2	RDD - mixed polynomial	0.0186	660	0.942	0.0016
3	RDD - quadratic	0.0203	710	0.930	0.0013
4	**RPJKD - linear**	0.0046	670	0.910	0.0022
5	RPJKD - quadratic	0.0059	710	0.985	0.0001
6	**RPJKD - mixed polynomial**	0.0039	550	0.996	0.0005
7	RPJKD - interacted quadratic	0.0056	720	0.993	0.0005
8	RKD - linear	0.0059	730	0.879	0.0031
9	RKD - quadratic	0.0166	730	0.910	0.0070
10	**RKD - mixed polynomial**	0.0043	730	1.000	0.0006
11	RKD - interacted quadratic	0.0067	730	0.997	0.0008
12	**birth cohort-IV - linear**	0.0042	670	0.912	0.0020
13	birth cohort-IV - quadratic	0.0056	720	0.998	-0.0003
14	birth cohort-IV - cubic	0.0060	700	0.990	-0.0004
WA	Inv-MSE weighted average	0.0039	As above	n.d	0.0010

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. The only difference is the set of bandwidths considered. The left side bandwidth is fixed at 365 days, the right side bandwidths considered range from 365 days to 730 days. As in Table 5, results are shown for the bandwidths which yield the lowest RMSE for each model. There are 1,461 treatment effect estimates for every model, one for each placebo-zone threshold. The smaller number of repetitions is a result of the larger maximum bandwidth considered.

The results for Version 2 of this exercise are summarized in Table 7. They are similar to those of the previous exercise – models 12, 10 and 6 continue to perform well. Optimal bandwidths vary, but are generally considerably larger than the baseline exercise. Model 12

has the lowest RMSE, with an optimal LHS bandwidth of 560 days. This is the single best performing specification amongst all candidates for estimating the effect of the 2007 reform.

Table 7: Candidate model performance in the placebo zone V2: Asymmetric bandwidths

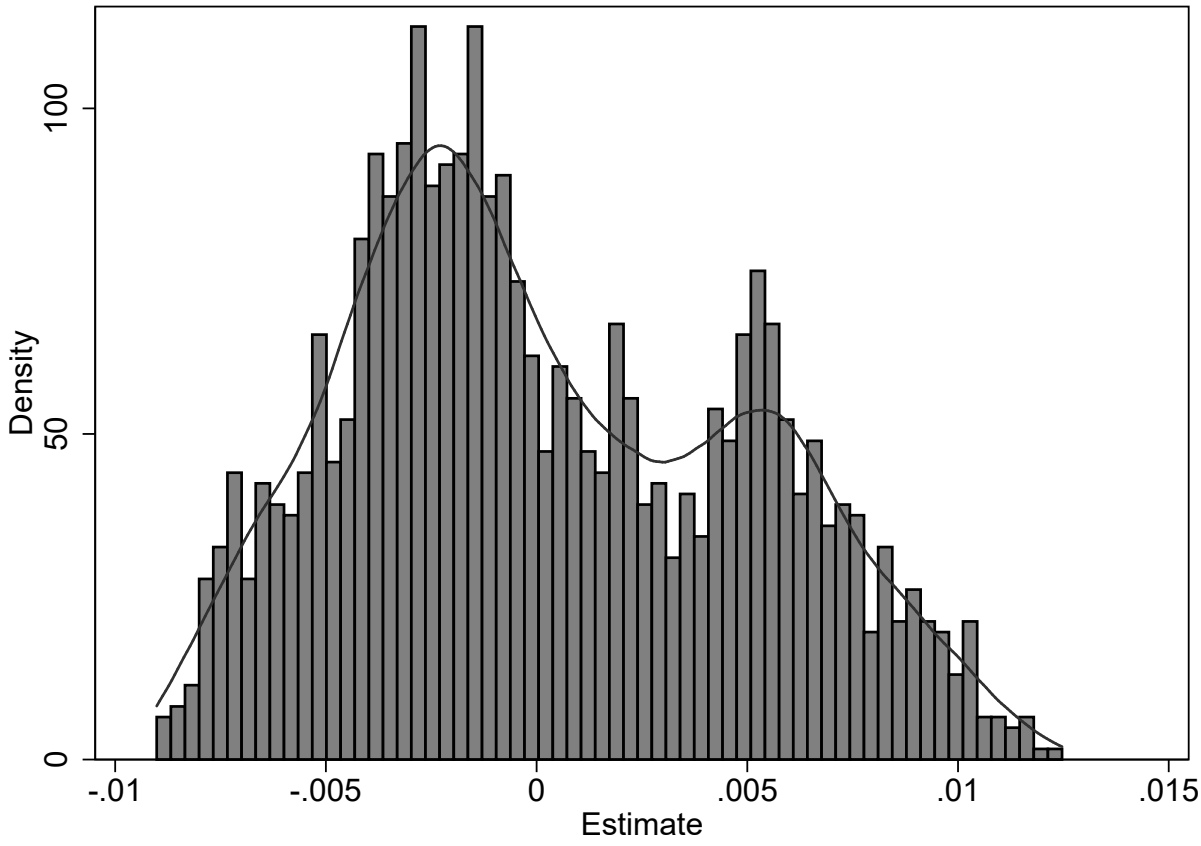
Model	Description	RMSE	Optimal LHS BW	Coverage	Bias
1	RDD - linear	0.0049	660	0.986	-0.0009
2	RDD - mixed polynomial	0.0095	730	0.979	0.0020
3	RDD - quadratic	0.0122	730	0.977	0.0019
4	RPJKD - linear	0.0047	690	0.987	-0.0005
5	RPJKD - quadratic	0.0041	610	0.999	-0.0009
6	**RPJKD - mixed polynomial**	0.0040	560	0.999	-0.0004
7	RPJKD - interacted quadratic	0.0122	730	0.969	-0.0001
8	RKD - linear	0.0051	370	1.000	-0.0002
9	RKD - quadratic	0.0054	600	0.990	-0.0012
10	**RKD - mixed polynomial**	0.0037	550	0.997	-0.0005
11	RKD - interacted quadratic	0.0186	370	0.942	0.0024
12	**birth cohort-IV - linear**	0.0036	560	1.000	-0.0004
13	**birth cohort-IV - quadratic**	0.0037	610	1.000	-0.0009
14	birth cohort-IV - cubic	0.0061	730	0.997	0.0007
WA	Inv-MSE weighted average	0.0038	As above	n.d	-0.0005

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. The only difference is the set of bandwidths considered. The right side bandwidth is fixed at 365 days, the left side bandwidths considered range from 365 days to 730 days. As in Table 5, results are shown for the bandwidths which yield the lowest RMSE for each model. There are 1,461 treatment effect estimates for every model, one for each placebo-zone threshold. The smaller number of repetitions is a result of the larger maximum bandwidth considered.

4.5 Using placebo zone estimates for inference

The placebo zone consists of 1,826 overlapping data windows, corresponding to 1,826 separate placebo estimates for each (symmetric) estimator. Consider the distribution of these placebo estimates for Model 12 – shown in Figure 4. One can use this distribution for alternative approaches to inference – randomization inference, in the spirit of Ganong and Jäger (2018). We discuss two alternative inference approaches. These alternatives may be useful if there is reason to believe that a given estimator or its estimated variance, are biased.

Figure 4: Distribution of placebo estimates from Model 12 (symmetric bandwidth)



Notes: This figure shows the distribution of treatment effect estimates generated within the placebo zone by 'Model 12', with a symmetric 365 day bandwidth. This is the model which performed best in the placebo zone trials reported in Table 5. Bars represent estimates grouped into 64 evenly sized bins. Kernel density fit is overlaid.

4.5.1 Approach 1: Fully non-parametric

Consider an estimate which lies outside of the range of the placebo estimates. If these 1,826 placebo estimates were independent, one would conclude that the two-sided p-value $< 2/1826 = 0.0011$. For an estimate lying inside the range of placebo estimates, $p = 2 * \min(i/1826, (1826 - i)/1826)$, where i is the rank of the estimate alongside the 1,826 placebo estimates.

However, the 1,826 placebo estimates are not independent. Indeed they are strongly serially correlated in our application. This is almost certainly the case in other applications as well, given the rolling data window. In our model 12, the serial correlation of placebo

estimates = 0.9895. This equates to an effective sample size (ESS) of just 10 independent observations.[19] A more appropriate two-sided p-value for estimates lying outside the placebo zone is $p < 2/ESS$. In our case, for model 12, $p < 2/10 = 0.2$.

4.5.2 Approach 2: Semi-parametric

A more powerful approach to inference is to calculate t-statistics, based on the distribution of placebo estimates, taking into account the effective sample size of those placebo estimates. This approach respects the fact that these placebo estimates are not independent, but invokes an assumption that they are drawn from a normal distribution. The mean of that normal distribution is not set to zero, but to the mean placebo estimate, thereby accounting for potential systematic bias. For example, for Model 12, the mean placebo estimate is 0.0002, with a standard deviation 0.0047. For a given estimate of the actual treatment effect $\hat{\beta}$, the t-stat = ($\hat{\beta}$ – (0.0002))/ 0.0047, distributed with ESS-1 degrees of freedom. The use of the t-statistic with ESS degrees of freedom takes into account the sampling error in the estimated variance of the population distribution of placebo estimates. For example, if $\hat{\beta}$ = 0.132 using model 12, t = -2.86, which corresponds to a p-value = 0.0187 assuming 9 degrees of freedom.

4.6 Treatment effect heterogeneity

The exercise above has allowed us to evaluate the performance of candidate models under a data generating process in which the treatment effect is precisely zero for all people. However, the results of this exercise might not be informative if treatment effects are heterogeneous. We now extend this exercise to more general data generating processes which incorporate treatment effect heterogeneity. We still use the placebo zone. This time we make modifications to the raw data to mimic key types of treatment effect heterogeneity.

To preview the results of this exercise, imposing purely random stochastic treatment

[19] This calculation draws on Eq. 5 in Zwiers and Storch (1995).

effects (Cases 1 and 2 below) does not change the conclusions of the exercise at all, it simply introduces noise. But imposing a non-constant marginal treatment effect (MTE) (Case 3) raises some interesting issues.

4.6.1 Case 1: Large stochastic treatment effects

In Case 1, we retain a zero mean treatment effect in the population. Noting the binary outcome variable (Y), and denoting treatment as T, we modify the raw data as follows:

- We stochastically impose a treatment effect of -1 for 'treated' people, whose $Y = 1$ in the original data. Their probability of being assigned this treatment effect is set to 0.02 divided by $E(Y|T = 1)$ in the original data. This would yield an expected ATE of -0.02 if this was the only change made. We chose -0.02 as this is close to (but larger than) our actual treatment effect estimates.
- We then assign an offsetting treatment effect of 1 for some 'treated' people whose $Y = 0$ in the original data. Again, this is done stochastically. The probability of being assigned this treatment effect is set so as to retain an expected treatment effect of zero within each date of birth.

Table 8 summarizes the performance of each model under imposing Case 1 treatment effect heterogeneity. Overall, the results are very similar to those in Table 5. Indeed the rank of each model (in terms of RMSE) is almost completely unchanged. The only notable difference is the larger RMSE for each model, which is an expected consequence of the additional stochastic variance that was imposed.

4.6.2 Case 2: Extreme non-monotonic stochastic treatment effects

Case 2 is similar to Case 1, but more extreme. Here, we impose a treatment effect of -1 for every treated observation whose $Y = 1$ in the original data. We then offset this with a stochastic treatment effect of 1 for some treated people for whom $Y = 0$. Similarly to Case

Table 8: Candidate model performance in the placebo zone: Heterogeneous treatment effects Case 1

Model	Description	RMSE	Optimal BW	Coverage	Bias
1	RDD - linear	0.0103	365	0.950	-0.0010
2	RDD - mixed polynomial	0.0228	365	0.929	0.0007
3	RDD - quadratic	0.0270	365	0.925	0.0011
4	**RPJKD - linear**	0.0072	365	0.942	0.0002
5	RPJKD - quadratic	0.0089	365	0.955	-0.0011
6	**RPJKD - mixed polynomial**	0.0067	365	0.972	-0.0007
7	RPJKD - interacted quadratic	0.0156	365	0.931	-0.0001
8	RKD - linear	0.0107	363	0.910	0.0019
9	RKD - quadratic	0.0203	365	0.939	0.0013
10	**RKD - mixed polynomial**	0.0071	365	0.972	-0.0006
11	RKD - interacted quadratic	0.0201	365	0.936	0.0013
12	**birth cohort-IV - linear**	0.0065	365	0.970	-0.0002
13	birth cohort-IV - quadratic	0.0085	365	0.957	-0.0013
14	birth cohort-IV - cubic	0.0143	365	0.926	-0.0005
WA	Weighted average - inv MSE	0.0069	365	n.d.	-0.0003

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. Here, however, random treatment effect heterogeneity has been imposed into the placebo zone, as detailed in the text.

1, the probability of being assigned this treatment effect is set so as to retain an expected average treatment effect of zero within each date of birth.

Table 9 summarizes the performance of each model under imposing Case 2 treatment effect heterogeneity. Again, the results are very similar to those in Table 5. The rank of each model (in terms of RMSE) is indeed unchanged. The RMSE for each model is slightly larger than in Case 1, as expected.

4.6.3 Case 3: Non-constant marginal treatment effect (MTE)

Case 3 is the most interesting and complicated of these three scenarios. Here, we impose an MTE which varies linearly with 'resistance' (see the discussion in Cornelissen et al., 2016, for a non-technical introduction to such models). The imposed MTE is -0.05 for those with lowest 'resistance' (resistance = 0). The imposed MTE is zero for those with highest resistance (resistance = 1). The MTE is set to be linear in resistance between these two

Table 9: Candidate model performance in the placebo zone: Heterogeneous treatment effects Case 2

Model	Description	RMSE	Optimal BW	Coverage	Bias
1	RDD - linear	0.0114	365	0.953	-0.0006
2	RDD - mixed polynomial	0.0249	365	0.926	0.0011
3	RDD - quadratic	0.0287	365	0.934	0.0015
4	**RPJKD - linear**	0.0082	365	0.943	0.0009
5	RPJKD - quadratic	0.0099	365	0.952	-0.0005
6	**RPJKD - mixed polynomial**	0.0077	365	0.961	-0.0001
7	RPJKD - interacted quadratic	0.0174	365	0.937	0.0006
8	RKD - linear	0.0117	360	0.923	0.0029
9	RKD - quadratic	0.0230	365	0.939	0.0024
10	**RKD - mixed polynomial**	0.0081	365	0.963	0.0002
11	RKD - interacted quadratic	0.0228	365	0.935	0.0024
12	**birth cohort-IV - linear**	0.0074	365	0.957	0.0005
13	birth cohort-IV - quadratic	0.0094	365	0.955	-0.0008
14	birth cohort-IV - cubic	0.0156	365	0.930	0.0003
WA	Weighted average - inv MSE	0.0079	365	n.d.	0.0004

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. Here, however, an extreme form of random treatment effect heterogeneity has been imposed into the placebo zone, as detailed in the text.

extremes. The intuition for such an MTE profile is in the spirit of a Roy Model. People with low resistance are those who may have the most to benefit from treatment. They select into treatment even if few otherwise similar people do. In other words, they are treated even when $E(T|Z)$ is low, where T denotes treatment and Z is an instrumental variable. Those with high resistance select into treatment only when $E(T|Z)$ is high.

Consider a relatively low-resistor, whose resistance is equal to say, 0.3. This person is induced into treatment by a marginal change in Z when $E(T|Z) = 0.3$. In our setting, these are people who are relatively slow to obtain a learners permit. It is their MTE that is identified by a marginal change in Z when $E(T|Z) = 0.3$. Similarly, consider a high-resistor, whose resistance is equal to say, 0.9. This person is induced into treatment by a marginal change in Z when $E(T) = 0.9$. In our setting, these are people who obtain a learners permit very soon after they are old enough to do so. It is their MTE that is identified by a marginal change in Z when $E(T|Z) = 0.9$.

In reality, the relationship between the MTE and 'resistance' may be completely different, and perhaps may have the opposite direction. People who obtain a learners permit as soon as they are old enough might actually have the most to gain from the treatment. This does not matter for our exercise, since our intention is simply to test the implications of an MTE which varies strongly.

Imposing a non-constant MTE highlights for the first time that the estimators are estimating different target parameters and hence need to be evaluated more subtly:

- RKD estimates MTEs at a specific threshold – which in our application is the extreme threshold with full resistance, where $E(T) = 1$, where we have imposed a MTE = 0.

- RDD estimates LATEs in a particular range of the MTE distribution. For example, for an RDD with an estimated first-stage discontinuity of 0.2, from $E(T) = 0.8$ to $E(T) = 1$, the LATE is the average of MTEs over this range. Given the linearity assumption, this equals the MTE at $E(T) = 0.9$, which is -0.005.

- The RPJKD estimates are identified by both the kink and the discontinuity. In the presence of treatment effect heterogeneity, the target parameter can be interpreted as a weighted average of the MTE (as estimated by the RKD estimator) and the LATE (as estimated by the RDD estimator).

- The month-of-birth cohort-IV strategy employs a vector of mutually exclusive instrumental variables. As discussed by Angrist and Pischke (2009, p. 174), it therefore identifies a weighted average of the LATEs identified by each individual month indicator variable. The weights are proportional to the inverse variance of each individual LATE estimate. Whilst easy to implement such a regression, and to estimate instrument-specific LATEs, it is not easy to pre-specify this weighted-average LATE, particularly given that the DOB-polynomial control variable is likely to have a large influence on each instrument's strength in the first-stage regression. However, in any month-of-birth IV regression without further controls, the first-stage predicted values

will always be exactly equal to the $E(T)$ within each month. The effect of each month-of-birth IV is therefore to induce treatment for people whose resistance is somewhere between $E(T|m1)$ and $E(T|m2)$, where $m1$ is the month with the lowest $E(T)$ and $m2$ is the month with highest $E(T)$. In light of this, we set the target parameter to equal the MTE where resistance equals the unconditional $E(T)$ – i.e. the mean treatment probability across the sample. For example, in the main estimation sample $E(T) = 0.79$. At this level of resistance, the imposed MTE = -0.0105. This is therefore an approximation of the target parameter for the month-of-birth cohort-IV models. However, $E(T)$ varies between placebo samples, and the target parameter is also allowed to vary in the placebo zone estimations.

Table 10 summarizes the performance of each model under imposing Case 3 treatment effect heterogeneity. This table has an additional Column "Target" – the target parameter, which as discussed above, differs between models. The target parameter for the RKD models is zero – the MTE at the threshold. The target parameter for the RDD models is the LATE that corresponds with the imposed MTEs through the estimated discontinuity. For RPJKD, we set the target to be the unweighted average of the MTE at the threshold, and the LATE associated with each discontinuity. For the cohort-IV, the target parameter is the MTE at the mean treatment probability through the sample, as discussed above.

The results in this table have some similarities and some differences to those in earlier tables. For most (but not all) estimators, a full 365 day bandwidth is preferred. Model 8 is the main exception. The coverage of the models is much more varied. In particular, coverage is poor for all of the RKD and the RPJKD models. These are also the models for which bias is relatively high.

Consistent with each other version, Model 12 (cohort-IV, controlling for a linear secular trend) has the lowest RMSE. Indeed, in this version, its RMSE is much lower than any other models. Model 13 (quadratic trend) now has the second lowest RMSE. The coverage rate for both of these models is also good. Amongst the other models, model 1 (RDD with linear

Table 10: Candidate model performance in the placebo zone: Heterogeneous treatment effects Case 3 – Incorporating non-constant MTE

Model	Description	RMSE	Optimal BW	Coverage	Target	Bias
1	**RDD - linear**	0.0087	365	0.961	-0.0068	0.0013
2	RDD - mixed polynomial	0.0209	365	0.915	-0.0047	0.0036
3	RDD - quadratic	0.0236	365	0.921	-0.0047	0.0037
4	RPJKD - linear	0.0121	363	0.447	-0.0034	-0.0106
5	**RPJKD - quadratic**	0.0091	365	0.836	-0.0034	-0.0049
6	RPJKD - mixed polynomial	0.0105	365	0.634	-0.0023	-0.0090
7	RPJKD - interacted quadratic	0.0160	365	0.809	-0.0023	-0.0083
8	RKD - linear	0.0216	260	0.558	0.0000	-0.0189
9	RKD - quadratic	0.0236	365	0.726	0.0000	-0.0152
10	RKD - mixed polynomial	0.0140	365	0.511	0.0000	-0.0126
11	RKD - interacted quadratic	0.0234	365	0.725	0.0000	-0.0152
12	**birth cohort-IV - linear**	0.0051	365	0.948	-0.0129	-0.0012
13	**birth cohort-IV - quadratic**	0.0084	365	0.931	-0.0129	0.0042
14	birth cohort-IV - cubic	0.0131	365	0.898	-0.0129	0.0033

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. Here, however, treatment effect heterogeneity has been imposed into the placebo zone. As detailed in the text, this heterogeneity incorporates a non-constant marginal treatment effect that is proportional to 'resistance'.

interacted trend) has the lowest RMSE and good coverage at 96%. Model 5 (RPJKD with a quadratic trend), has the next lowest RMSE, though its coverage rate is not as good.

The results in this section highlight that care must be taken to carefully consider the implications of heterogeneous MTEs in any given application. In our case, the placebo zone exercises overwhelmingly point to model 12 (with a full 365 day bandwidth) as the best estimator amongst all available options. This is perhaps a surprising result, given the apparent theoretical superiority of the RDD and RKD estimators. However, whilst violations of identifying assumptions leads to bias, the magnitude of such bias is usually difficult to ascertain *a priori*. As with all comparisons between alternate options, there is a trade-off between bias and variance. In our application at least, it seems that this trade-off is optimized by the between-cohort-IV estimator.

5 Results

The main estimation results are presented in Table 11. Panel A shows the estimated effects of the 2000 reform and Panel B shows the estimated effects of the 2007 reform. Each panel shows results from five separate estimators – one in each column.

Table 11: Main results

	Best estimator	Best symmetric estimator (cohort-IV)	Best symmetric RPJKD estimator	Best symmetric RKD estimator	Best symmetric RDD estimator
	(1)	(2)	(3)	(4)	(5)
	A: 2000 Reform (0 → 50 hours)				
MVA 1-year	-0.0144***	-0.0132***	-0.0147***	-0.0144**	-0.0168***
SE	0.0041	0.0049	0.0050	0.0058	0.0058
p-value	0.0005	0.0073	0.0032	0.0129	0.0038
alt. p-value	0.0101	0.0187	0.0161	0.0374	0.0578
Model	6	12	6	10	1
BW	365 / 550	365	365	365	365
	B: 2007 Reform (50 → 120 hours)				
MVA 1-year	0.0021	0.0003	0.0006	-0.0024	-0.0007
SE	0.0030	0.0033	0.0033	0.0046	0.0035
p-value	0.4790	0.9259	0.8477	0.6069	0.8422
alt. p-value	0.5532	0.9886	0.8882	0.6681	0.9524
Model	12	12	6	10	1
BW	560 / 365	365	365	365	365

Notes: This table shows the main estimated effects of the actual policy changes in our main application. Asymptotic standard errors are clustered at the DOB level. Alternate p-values use the randomization inference procedure described in Section 4.5. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Column (1) shows results from the 'best' estimator. This is the estimator with the lowest RMSE of all candidate models evaluated on the placebo zone. For the 2000 reform, this is 'Model 6', with a bandwidth of 365 days on the left and 550 on the right. For the 2007 reform, this is 'Model 12' with a bandwidth of 560 days on the left and 365 days on the right.[20]

[20]Recall that we are constrained to a maximum bandwidth of 365 days on the left of the threshold for the 2000 reform, and a maximum bandwidth of 365 days on the right of the threshold for the 2007 reform.

These 'best' estimates suggest that the first reform had a strong impact on reducing MVAs, while the second reform did not. The first reform is estimated to have reduced the crash rate by -0.014, a reduction of 21% relative to the predicted value at the threshold for the untreated. The conventional p-value associated with this estimate is 0.0005. We have no reason to be sceptical about the validity of this p-value, since this estimator was found to have good coverage in the placebo zone trial, as well as an estimated bias that is close to zero. Nevertheless, we also show alternate p-values, based on the distribution of placebo estimates, as discussed in the previous section. This p-value is larger (0.010), though still strongly significant. The alternate p-value is larger, primarily due to the small number of degrees of freedom used in the translation of the t-statistic into a p-value, which makes it inherently conservative.[21] For the 2007 reform, the alternate p-value is also slightly higher than the conventional p-value, but remains very far from any conventional threshold of statistical significance.

Column (2) shows results from the best symmetric estimator – which is Model 12 with a bandwidth of 365 days on each side. For the 2000 reform, all of the key parameters from this model are similar to those in Column 1. The standard error and both p-values are all slightly larger, but qualitatively the same as in Column (1). The estimate for the 2007 reform is close to zero.

Columns (3), (4) and (5) show the results from the best symmetric RPJKD, RKD and RDD estimators, respectively. In each case, the maximum feasible bandwidth (365 days) is used, consistent with the outcomes of the placebo zone trials. Again, the qualitative conclusions are the same, with strongly significant negative effects of the 2000 reform, and approximately zero for the second reform.

Table 12 delves deeper into the effects of the 2000 reform. Corresponding results for the

Model 6 is a RPJKD model with quadratic polynomials to the left and linear polynomial to the right of the threshold in each stage. Model 12 is a month-of-birth cohort-IV model, controlling for a linear secular relationship between DOB and the outcome variable.

[21] Just 6 degrees of freedom are used for this estimate. This is equal to the 'effective sample size' of placebo estimates calculated in the placebo zone minus 1, taking into account the very strong serial correlation of those estimates (0.9915).

2007 reform are generally precise zeros and are available on request. The structure of this table is the same as the previous table, and the same five estimators are used throughout.[22]

On possible explanation for the 2000 reform reducing MVA is delayed timing of obtaining a provisional license, which would support the idea that maturity rather than improved driving skill lowered MVAs. Appendix Figure A7 suggests a possible small delay effect. Panel A considers the extent to which this explains the main treatment effect. The first rows show the original estimates, while the next rows show estimates from the same models, but controlling for a quadratic of age (in days) of obtaining a provisional license. The estimated effects are generally slightly smaller when these controls are included. In the 'best' estimator, the treatment effect estimate is actually unchanged, while in the other models, this reduction is no more than 11%. Thus we conclude that delaying of obtaining a license is at most only a small factor in the treatment effects that we have estimated.[23]

Panel B shows results which consider the timing of the treatment effects. As may be expected, the majority (65% in the 'best' model) of the treatment effect is confined to the first 6 months after obtaining a provisional license. The effect in the 6-12 month period is also at least marginally significant across the estimators, and its magnitude is not small. The effect in the following year (12-24 months after obtaining a license) is not statistically significant in any column.

Panel C show results for serious MVAs. It shows strongly significant negative effects for the subset of MVAs in which one or more people were injured. The effect size (-0.0084 in the preferred model) is large (-30% relative to the predicted value at the threshold for the untreated). The estimate is larger when the other estimators are used. The effects for fatalities are not statistically significant, which reflects a lack of statistical power stemming from a relatively small number of fatalities.

[22]We use the same set of estimators across each of the outcome variables (and sub-populations) here. This approach has the advantage of transparency and internal consistency, which helps to interpret the drivers of the main estimates. An alternative approach is to choose a different set of preferred estimators (using the placebo zone approach) for each outcome variable and sub-population.

[23]Moreover, for the 2007 reform we observe a much stronger delay effect, yet our treatment effect estimates indicate no effect on MVAs.

Table 12: Further results for the 2000 reform

	Best estimator	Best symmetric estimator (cohort-IV)	Best symmetric RPJKD estimator	Best symmetric RKD estimator	Best symmetric RDD estimator
	(1)	(2)	(3)	(4)	(5)
	A: Age of Obtaining Provisional License (Mechanism)				
MVA 1-year	-0.0144***	-0.0132***	-0.0147***	-0.0144**	-0.0168***
SE	0.0041	0.0049	0.0050	0.0058	0.0058
controlling for age got P1s	-0.0144***	-0.0118**	-0.0133**	-0.0131**	-0.0155**
SE	0.0042	0.0052	0.0052	0.0061	0.0060
	B: Timing of Treatment Effect				
MVA 6 months	-0.0094***	-0.0074**	-0.0078**	-0.0072	-0.0093**
SE	0.0031	0.0035	0.0036	0.0044	0.0044
MVA 6-12 months	-0.0048*	-0.0059*	-0.0070**	-0.0069*	-0.0077*
SE	0.0028	0.0034	0.0034	0.0040	0.0040
MVA 1-2 years	0.0035	0.0026	0.0027	0.0019	0.0033
SE	0.0036	0.0045	0.0046	0.0053	0.0053
	C: Serious MVAs				
Injury	-0.0084***	-0.0093***	-0.0100***	-0.0102***	-0.0110***
SE	0.0026	0.0032	0.0032	0.0038	0.0036
Fatality	-0.0002	-0.0001	-0.0002	-0.0002	-0.0002
SE	0.0003	0.0004	0.0004	0.0005	0.0005
	D: Heterogeneity by Sex				
MVA 1-year males	-0.0132**	-0.0146**	-0.0139*	-0.0114	-0.0163*
SE	0.0059	0.0072	0.0073	0.0086	0.0085
MVA 1-year females	-0.0164***	-0.0111*	-0.0159**	-0.0181**	-0.0177**
SE	0.0056	0.0066	0.0068	0.0080	0.0083

Notes: This table shows further estimated effects of the first policy change in our main application. Asymptotic standard errors are clustered at the DOB level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Panel D shows results by sex. The preferred estimator suggests that the effects are similar by sex, as do the results of the other estimators.

5.1 Cost-benefit analysis

To better contextualize our results, we undertake a back-of-the envelope cost-benefit analysis. Expected social costs for non-injury MVAs are average property damage costs for motor vehicle accidents taken from BITRE (2009) ($4,004). For injury crashes, we take values from NRMA (2017) that are obtained using the willingness to pay method in respect of avoiding crashes with unknown injury ($144,172). Since our point estimates for fatalities are so imprecise as to be uninformative, we assume fatality risk decreased by the same percentage as injury risk in our baseline calculations. There are important caveats to this analysis, in particular i) our estimates are local average treatment effects and may lack external validity and ii) there are numerous ways of quantifying social costs, each with its own limitations. Further details and sensitivity analysis are in Appendix C.

Our estimates imply an average social gain of $2,300 per person due to the 50 MSDH reform. If we take the conservative view that on average people would complete 20 hours supervised in the absence of the reform, then this would constitute a net social improvement provided that supervisors' and learners' combined cost of obtaining hours is less than $46 per hour. Since we find no evidence the 120 MSDH reform improved safety, we cannot rule out nil social benefits for that reform.

6 Selected applications of the placebo zone approach

6.1 Head Start

To further demonstrate our approach, we now apply the placebo zone model selection algorithm to Ludwig and Miller (2007)'s RD analysis of the Head Start program. The data from this study have been used widely for illustrative purposes in the RD methodological literature, including papers by Calonico et al. (2014), Cattaneo et al. (2017), Ganong and Jäger (2018) and Calonico et al. (2019).

The unit of analysis is the county. Treatment is eligibility for technical assistance to develop Head Start funding applications. Eligibility is tied to a sharp, arbitrary cut-off in the county-level poverty rate at 59.198 percentage points. The outcome variable is child mortality from Head Start-relevant causes. Earlier papers have used a range of methodological approaches to estimate the same discontinuity. Ludwig and Miller (2007) prefer local-linear regressions with a triangular kernel. Citing a lack of consensus on bandwidth selection, they show results using bandwidths of 9, 18 and 36 percentage points, as well as from regular linear and quadratic specifications. Calonico et al. (2019) use the CCT bandwidth-selection algorithm, which yields bandwidths of 6.81 and 6.98, varying by the use of covariates.

We use the data from Calonico et al. (2019)'s replication files. We consider 10 separate estimators, each with a range of alternate bandwidths. These were chosen to examine questions of functional form (linear versus quadratic) kernel (uniform versus triangular), weights (unweighted or population-weighted), and covariates (include or exclude). *A priori*, weighted estimates are likely to be more precise, since residual variance is likely inversely proportional to population size, and population varies greatly (Mean = 38,964; Standard Deviation = 117,460), ranging from 224 to 2,664,438. We also show four estimates using CCT models. Models 11 and 12 are unweighted conventional and bias-corrected estimates with CCT bandwidths. Models 13 and 14 are corresponding weighted estimates.

We face two challenges for adopting our approach in this context. The first is a relatively small range of the forcing variable within the placebo zone. The placebo zone has a range of 47 percentage points (spanning 15.2 to 59.198 percentage points). When models with relatively large bandwidths are trialled, the effective sample size of the resulting placebo estimates is small. The second challenge is a considerably larger density (about 2.1 times larger) in the placebo zone than in the treatment zone (see Cattaneo et al., 2017, Figure A1). The results of trials within such a high density zone may not be relevant for choosing models to adopt in a low density zone. We address both of these challenges by splitting the

placebo zone sample into two independent groups.[24] We randomly allocated each county into one of these groups.[25] This solves the second challenge, since the resulting density is very similar to that of the treatment zone. It also helps with the first challenge, since the effective sample size of placebo estimates is approximately doubled.

The results of these placebo zone trials are shown in Table 13. For every model considered, the optimal bandwidths are either the maximum (15 percentage points), or close to it. This is considerably larger than the bandwidths in Calonico et al. (2019).[26] Since we are unable to test larger bandwidths, these should be seen as lower bounds for each optimal bandwidth. The results suggest that for this application, the population weights are very helpful – reducing the RMSE by around 28% in the linear model. The table also shows that covariates do not help, in fact they increase RMSE slightly. This is perhaps unsurprising, since the set of covariates is not rich and does not account for much residual variation.[27] The results suggest that models with a triangular kernel do worse than a regular rectangular kernel, and that a linear polynomial is preferred to higher orders. The CCT estimators (which are characterized by small bandwidths) perform poorly, but not to the same extent as they do in our main application.

The best performing estimator is the weighted linear RD, with no controls, and with full bandwidth. The estimated discontinuity using this estimator in the treatment zone is shown in Column (1) of Table 14. The estimate is statistically significant, consistent with Ludwig and Miller (2007) and with Calonico et al. (2019). But the estimate is also considerably

[24]We split the placebo zone observations into two groups because the placebo zone density is 2.1 times greater than the treatment zone density. This approach can be generalized for other contexts where the density is uneven. Practitioners may split the placebo zone into g groups, where g = round(placebo zone density / treatment zone density). It is not clear however if our approach is useful for situations where the treatment zone density is markedly greater than the placebo zone density.

[25]When these random allocations are repeated, the results are generally very similar. The RMSEs for the unweighted specifications are most sensitive to these repetitions, but they seem to always exceed the RMSEs for corresponding weighted specifications, usually by a large factor.

[26]The bandwidths in Calonico et al. (2019) are in turn similar to the average CCT-selected bandwidth within the placebo zone, which are shown in the last four rows of Table 13.

[27]The covariates are: percentage of black and urban population, levels and percentages of population in three age groups (children aged 3 to 5, children aged 14 to 17, and adults older than 25) as well as total population. We do not include 'total population' as a covariate whenever we use it as a weight instead.

Table 13: Head Start candidate model performance in the placebo zone

Model	Description	RMSE	Optimal LHS BW	Optimal RHS BW	Coverage	Bias
1	RD - linear	0.7763	15.0	15.0	0.972	-0.020
2	**RD - linear, weighted**	0.5616	15.0	15.0	0.844	-0.054
3	RD - linear, with covariates	0.7838	15.0	15.0	0.972	-0.019
4	RD - linear, weighted, with covariates	0.5622	15.0	15.0	0.876	-0.055
5	RD - linear, triangular kernel	1.0072	15.0	15.0	1.000	0.043
6	RD - linear, weighted, triangular kernel	0.6198	15.0	15.0	1.000	-0.036
7	RD - quadratic	1.4649	15.0	15.0	0.890	0.147
8	RD - quadratic, weighted	0.7751	14.6	14.6	0.918	0.017
9	RD - cubic	1.7051	14.6	15.0	0.968	0.254
10	RD - cubic, weighted	0.7842	15.0	15.0	0.982	0.057
C1	RD conventional	1.9513	4.3	4.3	0.954	0.078
C2	RD bias corrected	2.3386	4.3/6.9	4.3/6.9	0.961	0.052
C3	RD conventional - weighted	1.0278	5.1	5.1	0.972	0.089
C4	RD bias corrected - weighted	1.2098	5.1/8.0	5.1/8.0	0.968	0.084

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. However, these are for the Head Start application. The set of models considered is different, for reasons discussed in the text. The bandwidths considered ranged from 3 to 15 percentage points (in 0.2 percentage point increments) and was allowed to be asymmetric. There are 282 treatment effect estimates for every model, one for each placebo-zone threshold.

smaller than that of Calonico et al. (2019). The alternate p-value should be interpreted with some caution. It is relatively large primarily because the effective sample size from the placebo trial is small.

6.2 Minimum legal drinking age and drinking behavior

We now illustrate our approach with another application – discontinuities in drinking behaviour at the Minimum legal drinking age (MLDA). The MLDA context is one of the best known applications of RDD, beginning with Carpenter and Dobkin (2009). It is featured in econometric textbook treatments of RDD, such as Angrist and Pischke (2015).

Carpenter and Dobkin (2009) used restricted variables from the NHIS, which are not easily available. Instead, we draw on data from Lindo et al. (2016)'s corresponding analysis

Table 14: Head Start and MLDA RDD estimates

	Head Start	MLDA		
	Mortality	Ever Drinks	Drinks Regularly	Proportion of Days Drinks
	(1)	(2)	(3)	(4)
Estimated Effect	-1.323**	0.1816***	0.2344***	0.0712***
SE	0.5372	0.0236	0.0207	0.0065
p-value	0.0140	0.0000	0.0000	0.0000
alternate p-value	0.0756	0.0000	0.0000	0.0009
Model	2	3	3	3
BW	15.00	11.40	8.04	8.04

Notes: This table shows the main estimated effects for the Head Start and MLDA applicaitons, using the best-performing model (lowest RMSE) from the respective placebo-zone trials reported in Tables 13 and 15. Asymptotic standard errors are clustered at unique values of the running variable. Alternate p-values use the randomization inference procedure described in Section 4.5. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

for the Australian state of New South Wales. We use the same three self-reported drinking outcomes as Lindo et al.: 'Ever drinks', 'Drinks regularly' and 'Proportion of Days Drinks'. And we use the same data: waves 1-11 of the HILDA survey.

Following Carpenter and Dobkin (2009), Lindo et al. show estimates from linear specifications with bandwidths up to two years of age. These are centred around the 18th birthday MLDA threshold. Here, we consider the performance of a range of specifications – linear and quadratic, with and without weights, as well as CCT estimators. We consider a much wider bandwidth range, from three months to 12 years on the right side, with the left side capped at three years. The 3-year cap on the left reflects the limit of data availability in the treatment zone, since all respondents were aged 15 years and over.

Table 15 shows results from the placebo zone trials, for which the placebo zone consists of 18-50 year old respondents.[28] In many respects, the results are consistent across outcome variables used, and indeed consistent with the earlier applications we have shown: (i) long bandwidths are optimal for each estimator – much larger than those selected by CCT's

[28]The results are qualitatively similar when the placebo zone is changed (eg. 18-40 years, or 18-60 years) or if the maximum bandwidth is changed (e.g. 10 years, or 15 years). These are available on request.

procedure; (ii) linear RD yields the lowest RMSEs; (iii) weighting by cell-size reduces the RMSE; (iv) the CCT estimator does poorly, with or without bias adjustment.

Table 15: MLDA candidate model performance in the placebo zone

Model	Description	RMSE	Optimal RHS BW	Coverage	Bias
		A: Ever Drinks			
1	RD - linear	0.0196	5.41	0.961	-0.0007
2	RD - quadratic	0.0301	11.98	0.960	-0.0008
3	**RD - weighted linear**	0.0165	11.40	0.963	-0.0008
4	RD - weighted quadratic	0.0285	11.98	0.937	0.0002
C1	RD conventional	0.0572	1.02	0.908	-0.0003
C2	RD bias corrected	0.0659	1.02/1.60	0.908	-0.0009
		B: Drinks Regularly			
1	RD - linear	0.0272	8.12	0.977	0.0030
2	RD - quadratic	0.0406	9.92	0.981	0.0004
3	**RD - weighted linear**	0.0233	8.04	0.989	0.0012
4	RD - weighted quadratic	0.0357	10.42	0.979	-0.0002
C1	RD conventional	0.0632	1.07	0.976	0.0010
C2	RD bias corrected	0.0748	1.07/1.67	0.973	0.0010
		C: Proportion of Days Drinks			
1	RD - linear	0.0148	8.04	0.990	0.0003
2	RD - quadratic	0.0214	9.76	0.973	-0.0005
3	**RD - weighted linear**	0.0135	8.04	0.994	-0.0002
4	RD - weighted quadratic	0.0195	9.92	0.982	-0.0007
C1	RD conventional	0.0348	1.07	0.945	-0.0001
C2	RD bias corrected	0.0411	1.07/1.68	0.947	0.0000

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. However, these are for the MLDA application. The set of models considered is different, for reasons discussed in the text. For each of the three outcome variables, the bandwidth is allowed to vary from three months to 12 years on the right side, with the left side capped at three years. There are 622 treatment effect estimates for every model, one for each placebo-zone threshold.

Table 14 shows the estimated discontinuities at the MLDA, using the placebo-zone-optimal models we have identified. These are each weighted linear RD models, with a bandwidth of three years on the left, and between 8.04 and 11.40 years on the right, as per Table 15. These results are directly comparable to those in Lindo et al.'s Figure 3. Each of the point estimates is similar to Lindo et al.'s 2-year bandwidth estimates. But the standard errors are considerably smaller, suggesting a much narrower confidence interval for the effects of interest.

7 Conclusion and practical considerations

Regression Discontinuity Design and related estimators are amongst the most important tools of empirical economics. When using such estimators, however, applied researchers are typically faced with choosing between hundreds or thousands of candidate specifications. The large number of candidates is due to the numerous dimensions by which these estimators can vary – bandwidth, functional form, kernel, covariates are some of these dimensions, and these need not be the same on either side of the threshold. Various guidelines have been developed for model selection, but these generally only address one of these dimensions, whilst keeping others constant. In practice, contemporary applied work in leading economics journals still relies more on robustness testing than on model selection algorithms. Many such papers provide no explicit justification for model specification.

We have outlined a new approach for model selection which allows the performance of all candidate models to be assessed. The approach is conceptually straightforward. Each candidate model is assessed on its performance in estimating treatment effects in a placebo zone of the running variable – where the true effect is known to be zero. The RMSE of the resulting placebo estimates is the summary statistic by which each estimator is judged.

Our approach has potential to be useful for model-selection in a wide range of applications. We have demonstrated its use with three such applications within the paper. Researchers can implement the approach using our Stata command -pzms-.[29] However the approach should not be seen as a completely automated procedure for unproblematically choosing an objectively best specification. In this section we discuss some complications and suggestions for using the approach judiciously.

[29]We will release this program after the peer review process. In the meantime, if you would like to use our methods and need assistance, please contact us directly.

7.1 Estimators which rely on a 'first-stage'

Our method is relatively straightforward to apply for testing a set of candidate sharp-RD models. Other estimators (including RKD, fuzzy-RD, RPJKD, and cohort-IV) are defined with respect to a first-stage relationship between the running variable and the treatment variable. By definition, such a relationship does not exist in the placebo zone. With such models, we suggest the researcher imposes the actual first-stage relationship from the treatment zone into each iteration of the placebo-zone. This ensures that the source of identification is identical in the placebo zone estimates as it is in the treatment zone. Since there is no actual treatment, the assignment of a placebo treatment (in any way) has no implication for the integrity of the data generation process in the placebo zone.

In a practical sense, this means replacing the first-stage data around each placebo threshold with the first-stage data from the 'treatment zone', and then estimating the desired IV models. In many applications, this may require first collapsing the data to the level of the running variable. For example, in our application this means collapsing the data to the date-of-birth-level, relative to the threshold.

Our own application is unusual since the placebo treatments have a natural definition – as a function of date obtained learners license. This is the definition we have followed for the placebo zone testing throughout the paper. In Table 16 we show the results of the placebo zone testing when we instead impose the treatment-zone first-stage relationship as discussed above. As can be seen, the results of this exercise are quite similar to those from the original process. The same set of estimators performs best – although Model 6 now narrowly outperforms Model 12.

Table 16: Candidate model performance in the placebo zone: 'shifted' first-stage

Model	Description	RMSE	Optimal BW	Coverage	Bias
1	RDD - linear	0.0060	365	0.962	-0.0006
2	RDD - mixed polynomial	0.0114	365	0.922	-0.0001
3	RDD - quadratic	0.0129	270	0.971	0.0006
4	**RPJKD - linear**	0.0053	365	0.932	0.0006
5	RPJKD - quadratic	0.0057	365	0.978	-0.0004
6	**RPJKD - mixed polynomial**	0.0046	365	0.985	0.0000
7	RPJKD - interacted quadratic	0.0111	365	0.939	0.0007
8	RKD - linear	0.0113	350	0.921	0.0033
9	RKD - quadratic	0.0209	365	0.964	0.0012
10	**RKD - mixed polynomial**	0.0052	365	0.984	0.0001
11	RKD - interacted quadratic	0.0205	365	0.955	0.0013
12	**birth cohort-IV - linear**	0.0048	365	0.953	0.0004
13	birth cohort-IV - quadratic	0.0054	365	0.981	-0.0005
14	birth cohort-IV - cubic	0.0102	365	0.932	0.0004
WA	Inv-MSE weighted average	0.0050	365	n.d.	0.0001

Notes: The results in this table are from a similar procedure to what is detailed in the Table 5 notes. In this case, the first-stage relationship from the treatment zone is imposed into (each repetition of) the placebo zone, after collapsing the data to DOB level. This procedure is explained in the text.

7.2 How to set the maximum bandwidth for the placebo zone tests?

In any given application of our proposed method, the analyst must choose a maximum bandwidth for the set of candidate models. This choice will depend on the specific constraints of the application. In principle, one would like to consider all possible bandwidths, but this is not practical. If the chosen maximum bandwidth is too large, the number of thresholds within the placebo zone will be too small for the procedure to be informative about model performance.[30] We suggest that the effective sample size of placebo estimates should be taken into account when making this decision.[31] Given the large variation in model performance that we have observed in our applications, even relatively small effective sample sizes may be informative for model selection.

[30] Larger bandwidths are also likely to yield higher serial correlation in placebo estimates.

[31] We discuss the effective sample size of placebo estimates in Section 4.5.

7.3 Allowing for heterogeneous treatment effects

Our approach is perhaps most useful for model selection within (rather than between) a class of estimators. For example, consider the large set of candidate RDD estimators for a given application. Our approach assesses performance of such models with different bandwidths and different polynomial orders. Each of those candidate estimators has the same target parameter, and so comparing performance is relatively unproblematic.

Comparing performance between classes of models is more problematic, because they often estimate different parameters. Fuzzy-RDD models estimate LATEs, while RKDs estimate MTEs, RPJKDs estimate a weighted average of a LATE and a MTE (under additional assumptions of local MTE stability), while cohort-IV estimates a weighted average of a different set of LATEs. Our approach can be used to compare performance between such models. But this can only be done unproblematically if one is willing to assume that selection into treatment is unrelated to potential gains from that treatment. In our own application, this may be a reasonable assumption. It is less reasonable in many other applications.

More generally, researchers adopting our approach should carefully consider the implications of potential treatment effect heterogeneity. To be clear, placebo treatment effects in the raw data are precisely zero. This implies that model performance is assessed in a constant-treatment-effect context. This may be informative for model selection in more general contexts. But a more nuanced approach is to explore the implications for model performance if treatment effect heterogeneity is imposed into the placebo zone. The implications of a non-constant MTE may be particularly important to consider. In our application, imposing a non-constant MTE did not change the choice of 'best' overall model, nor within most classes of estimators. It did, however, raise doubts over the validity of the RKD and RPJKD estimators.

7.4 A partial test for plausibly-random treatment threshold

Our approach is informative for model selection only if the data generating process is similar in the treatment zone to that of the placebo zone. This is equivalent to assuming that the treatment threshold was effectively chosen at a random point of the combined support of the running variable.

This is difficult to test comprehensively. However, one particularly important facet of model selection is the extent of curvature in the conditional expectation function of the outcome variable (Y), with respect to the running variable (X). With more curvature, models with higher order polynomials and/or smaller bandwidths are likely to perform better.

In some cases, the treatment effect itself may confound the apparent extent of curvature in the raw data within the treatment zone. However, this is not the case if there is perfect compliance on one or both sides of the threshold. This includes the important case of Sharp RDD. For such cases, we propose the following test of curvature.

This simple test involves estimating a quadratic function on each side of the treatment zone (or just on one side if perfect-compliance is one-sided), and again separately, for every placebo zone section with a given bandwidth:

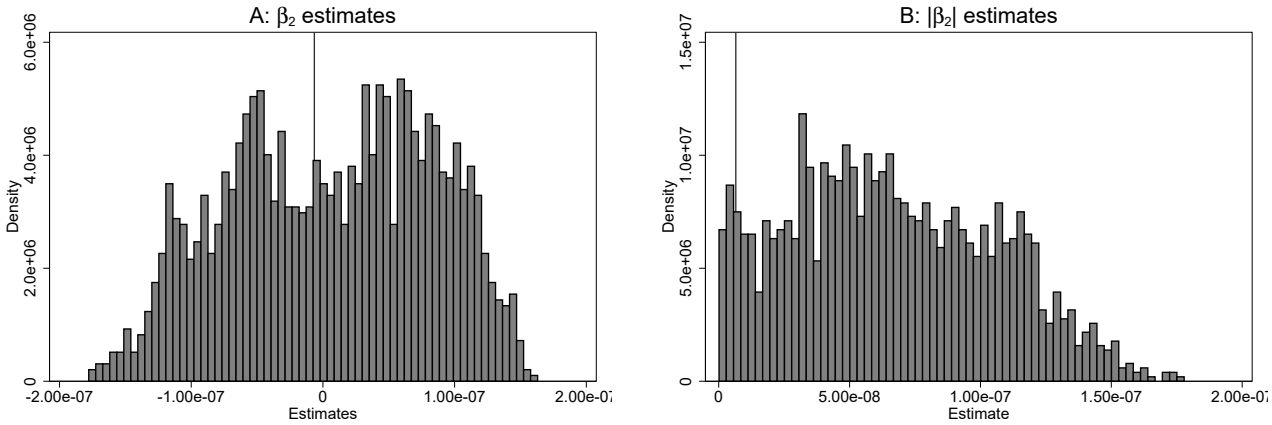
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u \tag{5}$$

The key parameter is $\hat{\beta}_2$. The proposed test is to consider whether $\hat{\beta}_2$ estimated in the treatment zone is atypical, as compared to those estimated in the placebo zone. This test uses the same randomization-inference approaches that we have discussed in Section 4.5. Perhaps more important, however, is to consider whether $|\hat{\beta}_2|$ is atypical, since the extent of curvature (as opposed to the direction of concavity) is more important for model selection.

The results of this exercise for our own application are conveyed succinctly in Figure 5. In our application, we have complete compliance on the RHS of the threshold, and so there is only one treatment-zone estimate for $\hat{\beta}_2$, denoted with the vertical line. Panel A shows

that $\hat{\beta}_2$ estimated in the treatment zone is near the middle of the distribution of placebo zone estimates. This supports the hypothesis that the treatment threshold was plausibly random chosen. Despite this, however, the extent of curvature in the treatment zone is very small, as compared to that of the placebo zone. This is shown in Panel B, which shows the $|\hat{\beta}_2|$ distribution.

Figure 5: Distribution of $\hat{\beta}_2$ estimates



Notes: This figure is a visual depiction of the proposed test of randomly chosen treatment threshold. The test seeks to determine whether the extent of curvature in the treatment zone is unusual. Panel A shows the distribution of $\hat{\beta}_2$, estimated through placebo zone, as per Eq. 5. Panel B shows the distribution of $|\hat{\beta}_2|$, which represents the extent of apparent curvature in the conditional expectation function within segments of the placebo zone. The corresponding estimates from the right side of the treatment zone are shown with vertical lines. Bars represent estimates grouped into 64 evenly sized bins corresponding to $\hat{\beta}_2$ values from estimating Eq. 5 throughout the placebo zone on 730-day segments. The vertical line corresponds to $\hat{\beta}_2$ ($|\hat{\beta}_2|$) estimated on the RHS of the 'treatment zone'.

What does this imply? As noted above, DGPs with relatively low curvature should have relatively low order polynomials and/or larger bandwidths. In our case, however, all of the best performing models already have a linear fit on the RHS and they already have the largest feasible bandwidth. We therefore conclude that our partial test of random threshold assignment provides no reason to depart from previous conclusions regarding model choice.

7.5 Randomization inference when the placebo zone is not contiguous

The placebo-zone approach facilitates an alternative approach to inference. We discuss this in Section 4.5, where we also emphasise the importance of taking account of serial correlation of the placebo estimates. In our own application, the placebo zone is contiguous. But in other applications it may not be, particularly if placebo data are available on 'both sides' of the real threshold. In such cases, it is not obvious how to determine the 'effective sample size', since the estimates on either side of the 'gap' may be correlated, but less so than estimates from immediately adjacent thresholds. One approach is to bound the effective sample size. The lower bound essentially ignores this discontinuity in serial correlation, and derives the ESS using a weighted average of the autocorrelations within each contiguous segment. The upper bound treats estimates between each segment as distinct and so the total ESS is the sum of the ESS in each segment.

In our application, this approach would produce tight bounds. To illustrate, if our placebo zone was not contiguous but instead consisted of two equally sized zones on either side of the treatment threshold, the lower bound for the ESS for our best symmetric bandwidth estimator would be 10 and the upper bound would be 11.

References

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, *114*, 533–575.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*.

Bates, L. (2012). *The experiences of learner drivers, provisional drivers and supervisors with graduated driver licensing in two Australian jurisdictions.* Queenland University of Technology: PhD Theis.

Bates, L., Watson, B., & King, M. (2010). Required hours of practice for learner drivers: A comparison between two Australian jurisdictions. *Journal of Safety Research*, *41*, 93–97.

Bates, L., Watson, B., & King, M. J. (2014). Parental perceptions of the learner driver log book system in two Australian states. *Traffic Injury Prevention*, *15*, 809–816.

Bates, L. J., Allen, S., Armstrong, K., Watson, B., King, M. J., & Davey, J. (2014). Graduated driver licensing: An international review. *Sultan Qaboos University Medical Journal*, *14*, e432–e441.

BITRE. (2009). *Cost of road crashes in Australia 2006.* Bureau of Infrastructure, Transport and Regional Economics: Research Report 118.

Bound, J., & Turner, S. E. (2002). Going to war and going to college: Did World War II and the G.I. Bill increase educational attainment for returning veterans? *Journal of Labor Economics*, *20*, 784–815.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, *101*, 442–451.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*, 2295–2326.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, *110*, 1753–1769.

Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, *83*, 2453–2483.

Card, D., Lee, D. S., Pei, Z., & Weber, A. (2017). Regression kink design: Theory and practice. In *Regression Discontinuity Designs (Advances in Econometrics, Vol. 38)* (pp. 341–382). Emerald Publishing Limited.

Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, *1*, 164–182.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations.* Cambridge University Press. doi: 10.1017/9781108684606

Cattaneo, M. D., Jansson, M., & Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *0*, 1–7.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2017). Comparing inference approaches for rd designs: A reexamination of the effect of Head Start on child mortality. *Journal of Policy Analysis and Management*, *36*, 643–681.

Chaplin, D. D., Cook, T. D., Zurovak, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, *39*, 403–429.

Chen, L.-H., Baker, S. P., & Li, G. (2006). Graduated driver licensing programs and fatal crashes of 16-year-old drivers: A national evaluation. *Pediatrics*, *118*, 56–62.

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, *41*, 47–60.

Cousley, A., Siminski, P., & Ville, S. (2017). The causal effects of World War II military service. *Journal of Economic History*, *77*, 838–865.

Dee, T. S., Grabowski, D. C., & Morrisey, M. A. (2005). Graduated driver licensing and teen traffic fatalities. *Journal of Health Economics*, *24*, 571–589.

Dong, Y. (2018). *Jump or kink? Regression probability jump and kink design for treatment effect evaluation.* Mimeo.

Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall.

Foss, R. D. (2007). Improving graduated driver licensing systems: A conceptual approach and its implications. *Journal of Safety Research*, *38*, 185–192.

Ganong, P., & Jäger, S. (2018). A permutation test for the regression kink design. *Journal of the American Statistical Association*, *113*, 494–504.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*, 447–456.

Gilpin, G. (2019). Teen driver licensure provisions, licensing, and vehicular fatalities. *Journal of Health Economics*, *66*, 54–70.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*, 201–209.

Hall, P. G., & Racine, J. S. (2015). Infinite order cross-validated local polynomial regression. *Journal of Econometrics*, *185*, 510–525.

Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O., & Tukiainen, J. (2018). When does regression discontinuity design work? Evidence from random election outcomes. *Quantitative Economics*, *9*, 1019–1051.

IIFHS. (2020, June). *Graduated licensing laws by state.* Insurance Institute for Highway Safety. Retrieved from `https://www.iihs.org/topics/teenagers/graduated-licensing-laws-table` (accessed 10 June 2020)

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, *79*, 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Imbens, G., & van der Klaauw, W. (1995). Evaluating the cost of conscription in the Netherlands. *Journal of Business and Economic Statistics*, *13*, 207–215.

Karaca-Mandic, P., & Ridgeway, G. (2010). Behavioral impact of graduated driver licensing on teenage driving risk and exposure. *Journal of Health Economics*, *29*, 48–61.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Perspectives*, *48*, 281–355.

Lindo, J., Siminski, P., & Yerokhin, O. (2016). Breaking the link between legal access to alcohol and motor vehicle accidents: Evidence from New South Wales. *Health Economics*, *25*, 908–928.

Ludwig, J., & Miller, D. L. (2005). *Does Head Start improve children's life chances? Evidence from a regression discontinuity design.* NBER Working Paper No. 11702.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*, 159–208.

Lyon, J. D., Pan, R., & Li, J. (2012). National evaluation of the effect of graduated driver licensing laws on teenager fatality and injury crashes. *Journal of Safety Research*, *43*, 29–37.

Masten, S. V., Foss, R. D., & Marshall, S. W. (2011). Graduated driver licensing and fatal crashes involving 16- to 19-year-old drivers. *JAMA*, *306*, 1098–1103.

McCartt, A. T., Teoh, E. R., Fields, M., Braitman, K. A., & Hellinga, L. A. (2010). Graduated licensing laws and fatal crashes of teenage drivers: A national study. *Traffic Injury Prevention*, *11*, 240–248.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*, 698–714.

McKnight, A. J., & Peck, R. C. (2002). Graduated driver licensing: what works? *Injury Prevention*, *8 (Suppl II)*, ii32–ii38.

Moore, T. J., & Morris, T. (2020, February). *Shaping the habits of young drivers.* Mimeo.

NRMA. (2017). *The cost of crashes: An analysis of lives lost and injuries on NSW roads.* National Roads and Motorists' Association.

O'Brien, N. P., Foss, R. D., Goodwin, A. H., & Masten, S. V. (2013). Supervised hours requirements in graduated driver licensing: Effectiveness and parental awareness. *Accident Analysis and Prevention*, *50*, 330–335.

Pei, Z., Lee, D. S., Card, D., & Weber, A. (2020). *Local polynomial order in regression discontinuity designs.* NBER Working Paper 27424.

Scott-Parker, B. J., Bates, L., Watson, B. C., King, M. J., & Hyde, M. K. (2011). The impact of changes to the graduated driver licensing program in Queensland, Australia on the experiences of Learner drivers. *Accident Analysis and Prevention*, *43*, 1301–1308.

Shope, J. T. (2007). Graduated driver licensing: Review of evaluation results since 2002. *Journal of Safety Research*, *38*, 165–175.

Steadman, M., Bush, J. K., Thygerson, S. M., & Barnes, M. D. (2014). Graduated driver licensing provisions: An analysis of state policies and what works. *Traffic Injury Prevention*, *15*, 343–348.

Traynor, T. L. (2009). The impact of state level behavioral regulations on traffic fatality rates. *Journal of Safety Research*, *40*, 421–426.

Trempel, R. E. (2009). *Graduated driver licensing laws and insurance collision claim frequencies of teenage drivers.* Highway Loss Data Institute.

WHO. (2018). *Global Status Report of Road Safety.* Geneva: World Health Organization.

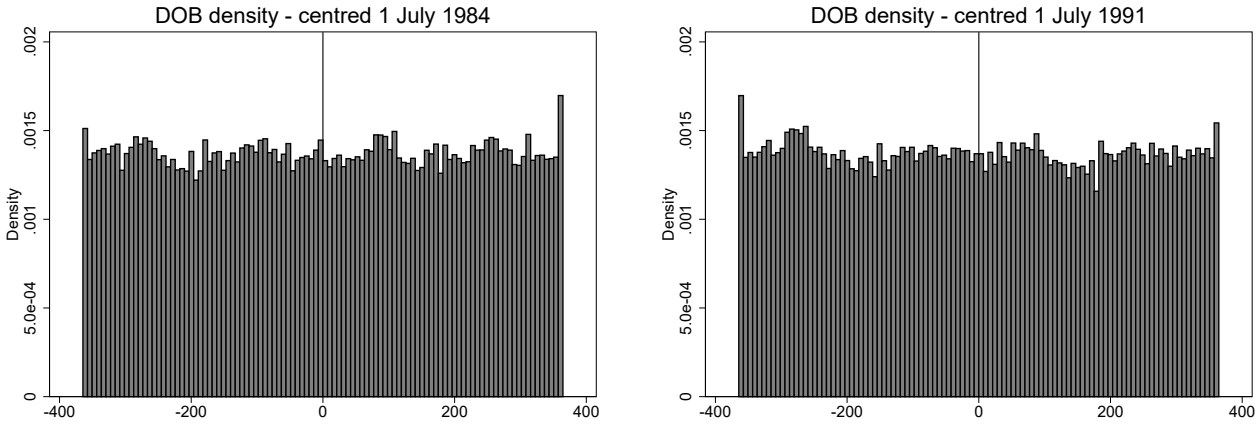
Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, *32*, 853–877.

Zwiers, F. W., & Storch, H. v. (1995). Taking serial correlation into account in tests of the mean. *Journal of Climate*, *8*, 336–351.

Online Appendices

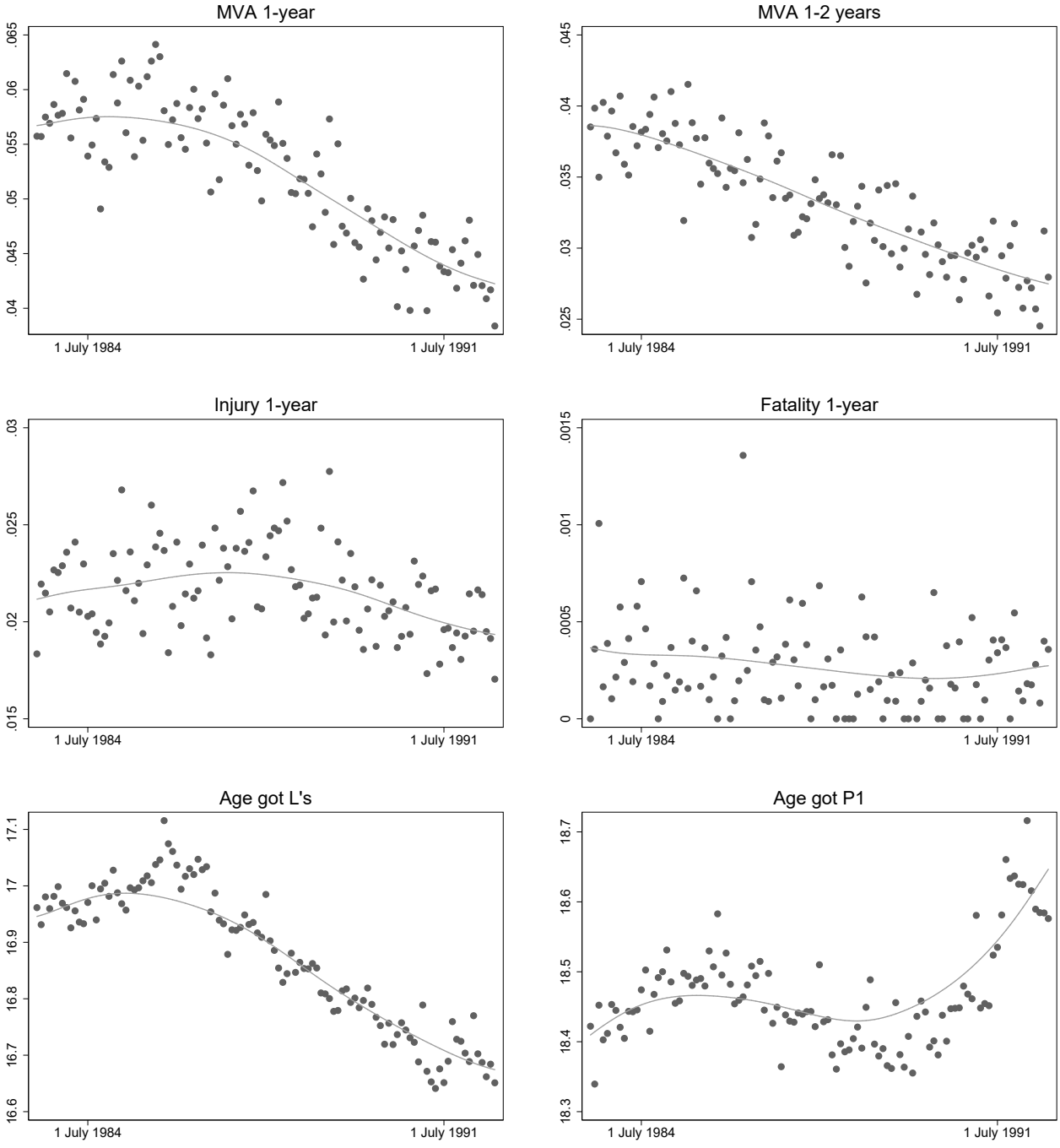
A Additional tables and figures

Figure A1: DOB distribution plots



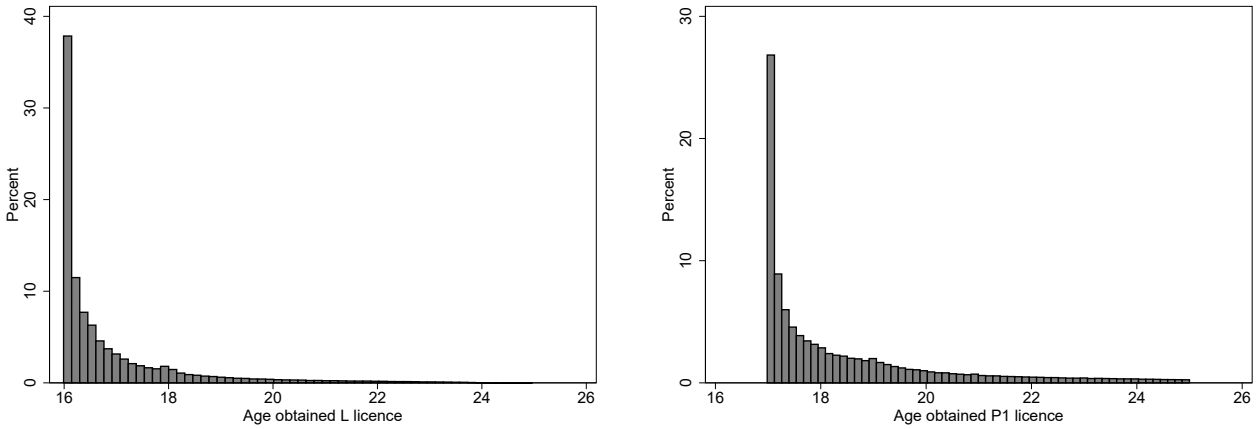
n = 154,524 in Panel A. n = 160,301 in Panel B. 7-day bins.

Figure A2: Scatter and fit plots: Main variables by DOB



Notes: Scatter plots 30-day bin size with lowess trend lines.

Figure A3: Distributions for age obtaining license



Notes: Sample includes all NSW licensees born between 1 July 1983 and 30 June 1992 ($n = 704,468$).

Figure A4: First-stage relationship between DOB and 120 MSDH treatment

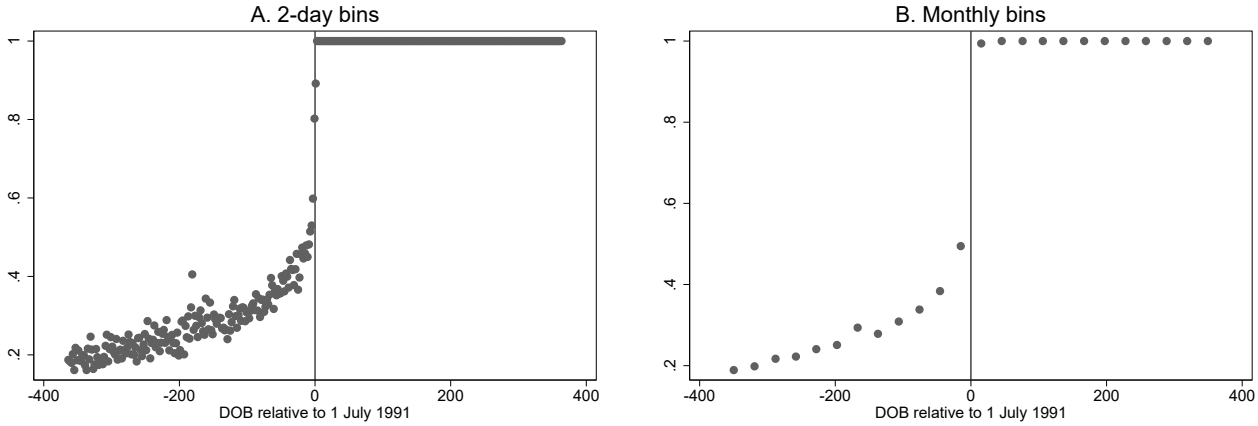
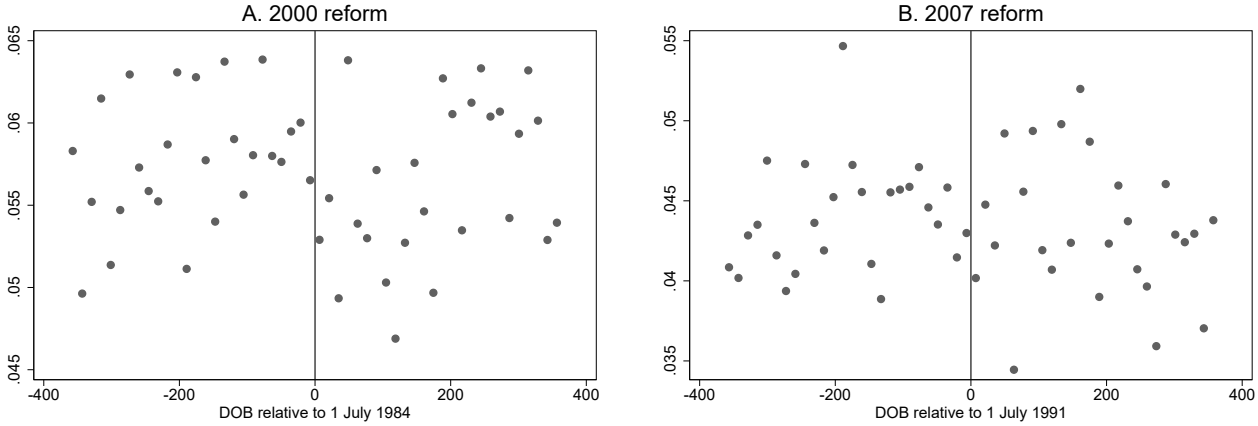
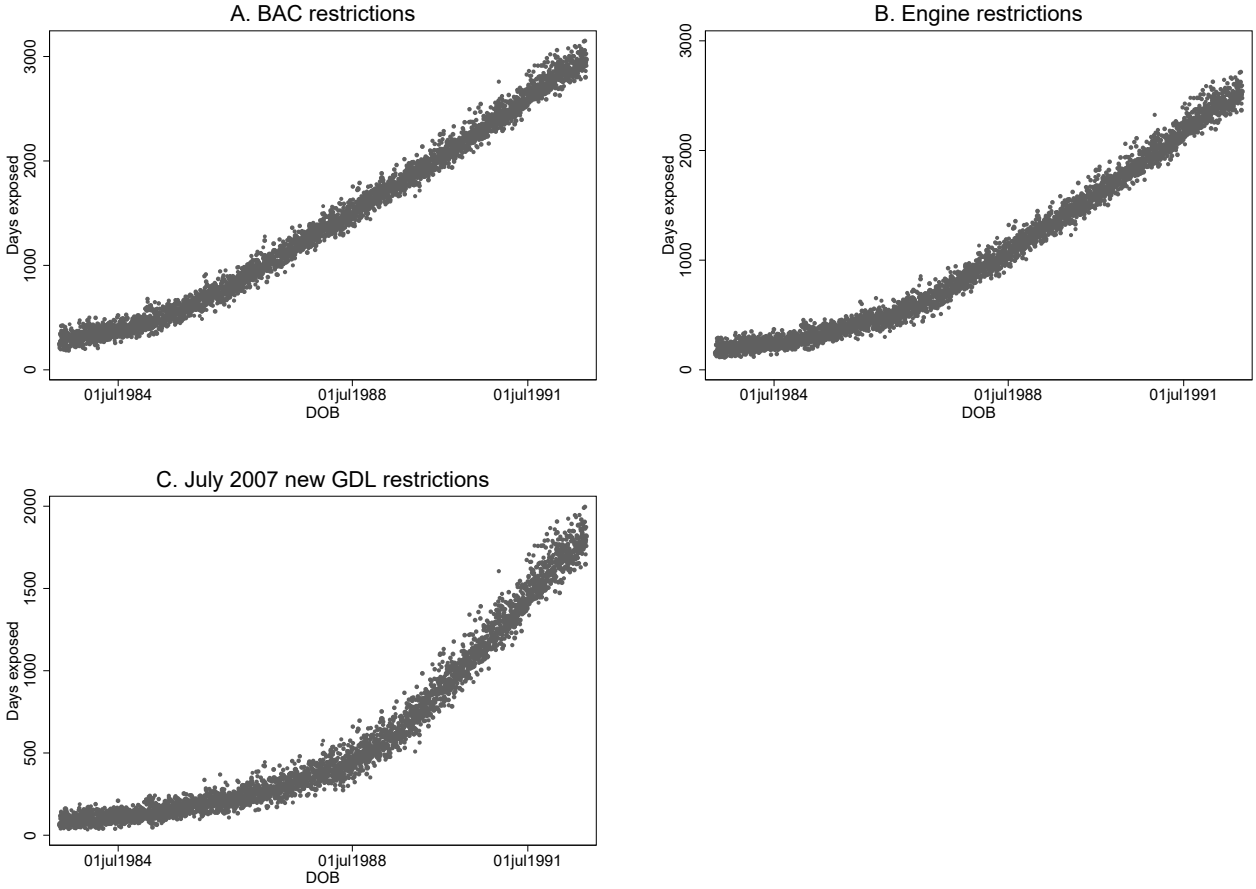


Figure A5: Reduced-Form Relationships between DOB and MVA 1-year



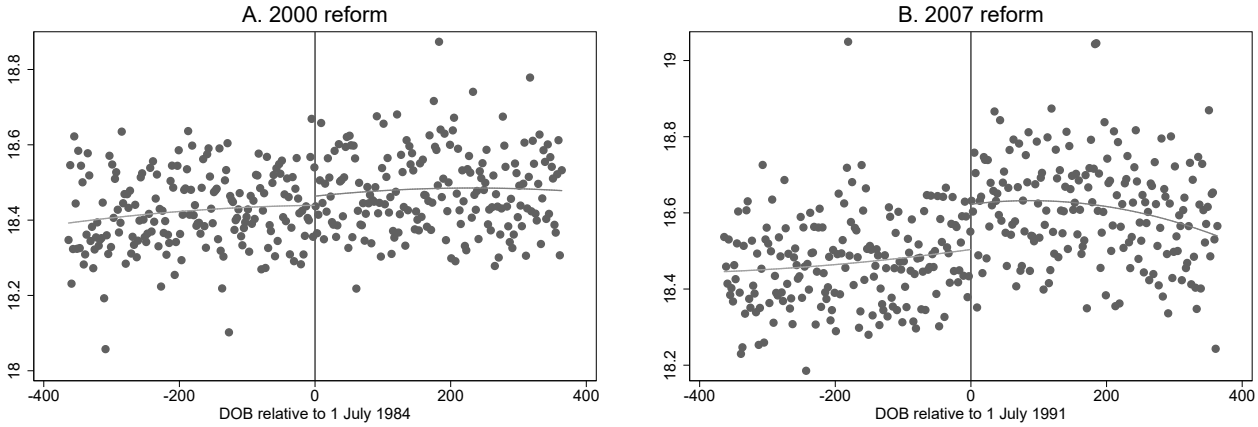
Notes: Scatter plots use 14-day bin size.

Figure A6: Exposure to BAC, engine and passenger restrictions by DOB



Notes: Day exposed is the maximum of zero and the date their P1 license expired minus the relevant policy change date.

Figure A7: Relationship between DOB and age obtained P1 license



Notes: Scatter point correspond to 2-day bins.

B Further details on the MVA and license datasets

B.1 Matching MVAs to driver license records

From October 2002 onwards, we can match 99.6% of MVAs involving drivers who, according to the MVA data are between 17-20 years old and licensed in NSW, to the license data.[32] Prior to this date the match rate is discontinuously lower by around 20 percentage points between July 2002 and October 2002 and 15 percentage points pre-July 2002 (see Appendix Figure B1). CRS cited an improvement in record keeping practices as a reason for the discontinuity but were unable to provide further details. Our analysis in Appendix Figure B1 indicates that the discontinuities are not limited to any subset of MVAs by characteristics, which would have allowed us to exclude inconsistently recorded MVAs. To address the missing MVAs we therefore inflate the MVA indicators we use as dependent variables for people who are not matched to an MVA (i.e. are recorded as having not had an MVA in our raw data) by a factor equal to the probability they actually did have an MVA given what we know about the rate of non-matched MVAs.

Focusing on our main dependent variable (any MVA within 12 months of obtaining P1 license), our preferred approach adjusts the crash probability for a person obtaining their P1 license on day t by:

$$1-\left(1-\left[\sum_{t}^{t+365}\frac{MVA_t}{n_t}\right]\times 0.15\right)^{\min\{t^*-t,365\}}\times\left(1-\left[\sum_{t}^{t+365}\frac{MVA_t}{n_t}\right]\times 0.20\right)^{\mathbf{1}.[t>t^*]\times\min\{t-t^*,130\}} \tag{B1}$$

where MVA_t is total number of matched MVAs involving drivers aged 17-20 years, n_t is the total number of licensed drivers aged 17-20 years (so that $\sum_t^{t+365}\frac{MVA_t}{n_t}$ is the unadjusted probability of being involved in an MVA within one year of obtaining a P1 license on day t), t^* indicates 1 July 2002 and $\mathbf{1}.[t>t^*]$ is an indicator for $t>t^*$. We also use Eq. B1 to adjust indicators for MVAs involving injury since the match rates for these crashes is almost identical to the rate for MVAs overall. For MVA indicators over six-month windows we replace 365 with 183.

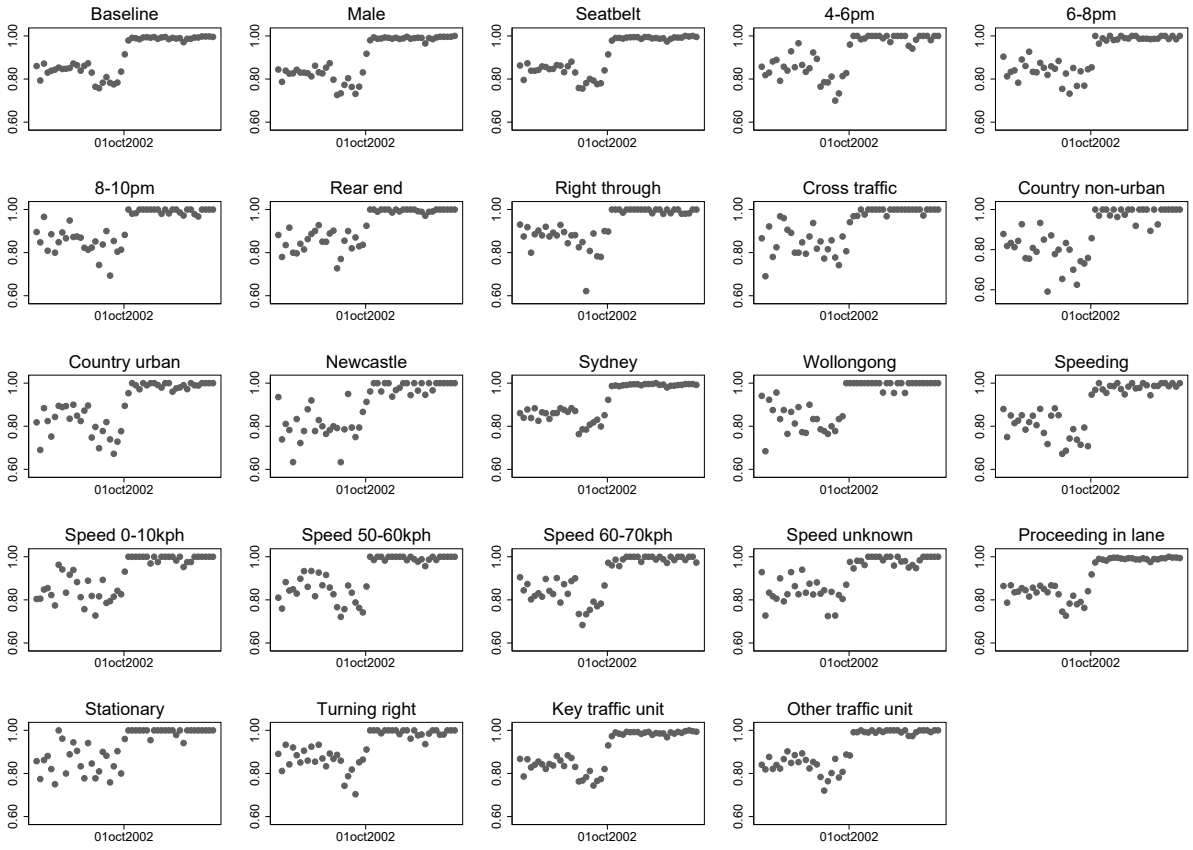
A more sophisticated approach of obtaining the adjustment factor is to replace 0.15 and 0.20, which are approximate rates at which MVAs can be matched to license data in the pre-July 2002 and July-October 2002 periods, with daily estimates for this rate and calculate

$$1-\sum_{t}^{t+365}\left(1-\left[\sum_{t}^{t+365}\frac{MVA_t}{n_t}\right]\times Match\ rate_t\right). \tag{B2}$$

This approach allows for more variation in the match rate; however, it may suffer from rare events bias since there are often few MVAs on a given day (expanding the time unit can solve this but setting a new time unit is arbitrary). In practice, the two approaches give very similar adjustment factors (Appendix Figure B2) and after confirming the choice

[32]The match rate is almost identical (99.5%) if we instead look at all people who, based on their age recorded in the MVA data, we can be certain were born after 1980.

Figure B1: Match rates: License and MVA data



Notes: Each scatter point corresponds to the average percentage of MVAs for 17-20 year old drivers that can be matched to license records for NSW licensed drivers (15 day groupings). Baseline: The full sample of 17-20 year old drivers; Male: Males only; Seatbelt: driver wearing seatbelt; 4-6pm: MVA between 4-6pm; 6-8pm: MVA between 6-8pm; 8-10pm: MVA between 8-10pm; Rear end: MVA reason, rear-ender; MVA reason, right through; Cross traffic: MVA reason, cross traffic; Country non-urban: MVA in country non-urban region; Country urban: MVA in country urban region; Newcastle: MVA in Newcastle region; Sydney: MVA in Sydney region; Wollongong: MVA in Wollongong region; Speeding: speeding involved in MVA; Speed 0-10kph: main vehicle travelling between 0-10kph; Speed 50-60kph: main vehicle travelling between 50-60kph; Speed 60-70kph: main vehicle travelling between 60-70kph; Speed unknown: main vehicle speed unknown; Proceeding in lane: manoeuvre before crash, proceeding in lane; Stationary: manoeuvre before crash, stationary; Turning right: manoeuvre before crash, turning right; Key traffic unit: vehicle was key traffic unit in MVA; Other traffic unit: vehicle was not key traffic unit in MVA.

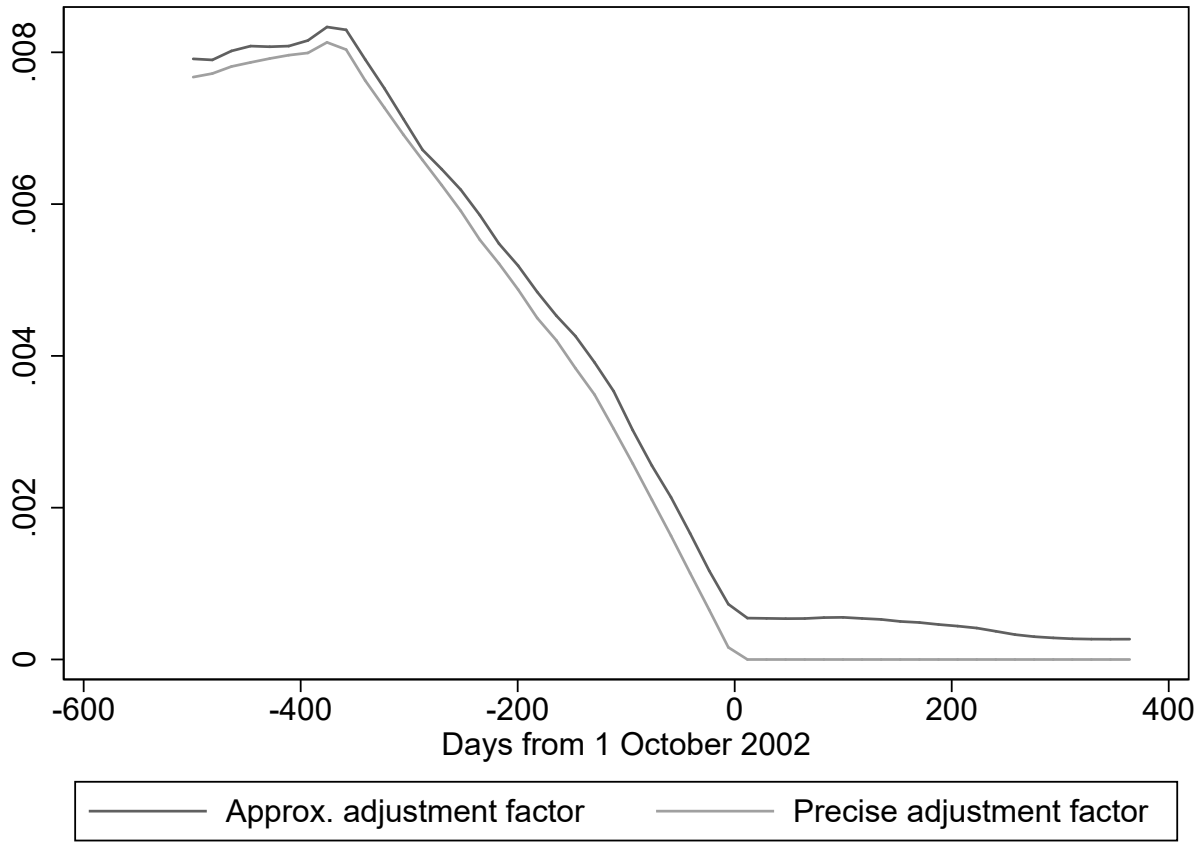
had no effect on our main results, we committed to using the simpler and more transparent approximation approach.

B.2 Sample restrictions

Throughout our analysis we always maintain the following sample restrictions:

- *Exclusion of people whose license history violates the GDL rules* – some people in our

Figure B2: Adjustment factors



Notes: The y-axis shows by how much the MVA within first year of P1 license probability is adjusted for a person obtaining their P1 license on day t (the x-axis) in order to account for missing MVA data. Formulaic details are in Eq. B1 (Approx. adjustment factor) and Eq. B2 (Precise adjustment factor).

dataset appear to have licencing histories that violate rules of the GDL system. For example, some people obtain their P1 license before turning 17 years; others are on their learner license for less than the mandatory period. These violations may have a variety of unobserved causes, such as people moving from interstate or data error. Only around 0.3% of people in our analysis sample violate the GDL rules and we exclude them throughout the analysis.

- *Exclusion of people whose eligibility to avoid treatment status is uncertain* – in our models, people cannot avoid 'treatment' if they are born after a certain date. People whose minimum and maximum possible DOB straddles the threshold date are dropped in the models we estimate (see discussion in Section 3). Recall that for people with multiple license records we can narrow this range. However, since the reforms may have affected licensing behavior (and consequently the accuracy of DOB), we only use the first license record when imposing this restriction.

- *Exclusion of people who obtained their P1 license after age 25* – these people are not

subject to the MSDH requirements. Only 8.7% of people in our dataset obtained their P1 license after age 25.

B.3 Additional details on the data

- *Change in MVA record keeping 2014* – in 2014, a policy change meant that NSW Police reported fewer MVAs from this year onwards. This was due to NSW Police no longer being refquired to attend a crash scene and investigate for tow away MVAs where nobody was injured or killed. This policy change is largely inconsequential for us as we only consider the periods July 2000-June 2008 in our analysis, and most people we observe obtain their P1 license more than 12 months before 2014. Moreover, exposure to this period (in terms of days on P1 after 2014) is a smooth function of DOB. We therefore ignore this change.

- *License suspensions and demotions* – in our analysis we focus on the first date a person obtained their P1 license and ignore any suspensions (e.g. for speeding or drunk driving), demotions (i.e. being made to redo the learner class due to a serious driving offence) or moves out of NSW after this date. If these events are unrelated to MSDH, which seems reasonable, then our MVA rates will be equally affected by them in the treatment and untreated groups. While we do not observe suspensions or moves in our data, we note that in 99.1% of cases the expiry date for learner license matches the date of effect for the first time obtaining a P1 license, which indicates demotions are rare.

C Cost-benefit analysis

Our cost-benefit analysis estimates the social benefits from reducing the probability of an MVA in the first 12 months of unsupervised driving. Since we find no evidence of reduced risk beyond this period, we assume those benefits are zero.

We proceed by first estimating the reduction in the rate of MVAs at the policy threshold for each MVA type (non-injury, injury and fatality) using our 'best' model for each reform (see Table 11). Since our point estimates for fatalities are imprecise owing to low frequencies, we assume fatality risk decreased by the same percentage as injury risk in our baseline calculations. We also show how the estimates change if instead we assume no change in the fatality risk.

We multiply the MVA rates by the total social costs associated with each type of MVA under the assumption that no one is treated ($T = 0$) and under the assumption that everyone is treated ($T = 1$). The difference between these estimates is the total social benefit per person. Social benefits for non-injury MVAs are taken from BITRE (2009). Specifically, we use the average repair cost for MVAs. We therefore ignore other costs associated with these MVAs such as towing costs, time lost and administrative fees associated with, for example, insurance claims. On the other hand, the BITRE value includes all MVAs, even severe MVAs resulting in injury or fatality, and as such the repair costs may be overstated. Moreover, non-injury crashes comprise a negligible proportion of total social benefits so doubling or tripling this value has little material effect on the estimates.

Social benefit estimates for crashes involving injury and fatality are taken from NRMA (2017). They are estimated using the willingness-to-pay method, which uses hypothetical scenario analysis to infer preferences. A strength of this approach is that it should capture all the information that goes into people's individual preferences. A drawback is that people may be unsure of their preferences, particularly if they have never experienced an MVA. This approach also ignores externalities, and may be subject to hypothetical bias.

Table C1 steps through our calculations for the 2000 (0 to 50 MSDH) reform. We have set out this table in such a way that it is easy to substitute our chosen social benefits parameters other parameters, if desired.

Our preferred estimate for the average social benefit is the sum of the average social benefit due to reduced risk of non-injury MVAs ($23), injury MVAs ($1,212) and fatalities, assuming that this risk falls by the same percentage as injury MVAs ($751). This implies an average social benefit of $2,300. If we ignore the reduction in fatalities, we estimate an average social benefit of $1,235.

Under the assumption that on average learners complete 20 hours before the reform and 50 hours after the reform, our estimates imply a social benefit of between $25-$46 per hour. By way of comparison, the national minimum wage in 2019-2020 is $19.49. Given that some supervised driving hours will be for trips that would have been taken anyway, and that there may be positive externalities to supervision (e.g. bonding), it seems likely, based on our estimates, that the 2000 reform was welfare improving.

Table C1: CBA – 2000 reform (0 to 50 MSDH)

	Non-injury	Injury	Fatality	Alt. fatality
Policy effect	-0.0058238	-0.008405	-0.0002131	-0.000101968
Prediction $T = 0$	0.039571	0.0279324	0.000480	0.000480
Prediction $T = 1$	0.0337472	0.0195274	0.0002669	0.000378
% reduction	-15%	-30%	-44%	-21%
Average social cost	\$4,004	\$144,172	\$7,369,845	\$7,369,845
Expected cost of driving $T = 0$	\$158.44	\$4,027.06	\$3,537.53	\$3,537.53
Expected cost of driving $T = 1$	\$135.12	\$2,815.30	\$1,967.01	\$2,786.04
Social saving per person	\$23.32	\$1,211.76	\$1,570.51	\$751.49

Notes: Policy effect estimates correspond to the 'best' estimator in Table 11. T means 'treated' (i.e. subject to the 50 MSDH policy). Expected cost of driving equals the predicted MVA probability $\times$ the average social cost. All values are expressed in 2019 \$AUD (for references, in 2019 the \$AUD:\$USD exchange rate averaged around 0.7:1).

D Summary of discontinuity studies

D.1 Literature search

We searched Econlit on 30 April 2020 using the terms "discontinuit*", "fuzzy RD" and "regression kink" (contained anywhere) and restricted results to papers published in 2019 in the following journals: *AEJ: Applied Economics*; *AEJ: Economic Policy*; *American Economic Review*; *Journal of Health Economics*; *Journal of Human Resources*; *Journal of Labour Economics*; *Journal of Political Economy*; *Journal of Public Economics*; *Quarterly Journal of Economics*; *Review of Economic Studies*; and *Review of Economics and Statistics*.

We identified 34 papers. Four papers were omitted because they did not use a discontinuity design, three were omitted because they were econometric theory focused rather than applied, and one paper was omitted because it was a reprint from 2018. This left us with 26 papers.

D.2 Explanation for columns in Table D1

We summarize the 26 papers in Table D1. Naturally, it was challenging to categorize these papers because they often used a variety of dependent variables and considered many different specifications, and it was often difficult to determine what the preferred model was. The Table reflects our own best judgement. Here we provide additional details to interpret columns that are particularly subject to our own judgements.

- *RD main specification?* – If the discontinuity design was a robustness check or used only as supporting evidence we categorized the paper as 'No'. For papers that used multiple estimation strategies, we coded them as 'Yes' provided the discontinuity estimates received (in our view) at least equal weight in the paper's conclusions to the other estimates.

- *Main model* – These categories are generally uncontroversial. For less-standard designs, we have adopted the descriptions used by the authors.

- *Main function* – In many cases, authors estimate the model using e.g. local linear regression, and then alter this specification in a sensitivity analysis section. In those cases, we would classify the main model as 'local linear'. In other cases, authors present a suite of models in a single table. Here we tried to classify the main model as the one that received the most emphasis when discussing results, unless a particular model was explicitly identified as being the main model or baseline specification.

- *How bandwidth?* – The categories are as follows. 'ROT' means a rule-of-thumb formulaic approach, such as Fan and Gijbels (1996). 'IK' refers to the approach in Imbens and Kalyanaraman (2012). 'CCT' refers to the approach in Calonico et al. (2014). 'Not discussed' means that the authors did not discuss bandwidth choice *a priori* to estimating their baseline results. This does not mean the authors did not consider bandwidth choice at all. In fact, in almost every paper the authors varied the bandwidth as a robustness exercise. Often, IK and CCT were used as robustness checks.

- How polynomial? – The categories are as follows. 'GI argument' means they chose a low-order polynomial based on the advice in Gelman and Imbens (2019). 'Cross-validation' means they used a cross validation procedure (i.e. curve fitting). 'Visual' means they motivated the choice based on visual evidence for the DGP. 'Significance tests' means they added higher order terms and tested whether they were statistically significant. 'AIC' means they chose the polynomial order with the lowest Akaike information criterion. Again, 'Not discussed' means that the authors did not discuss polynomial choice *a priori* to estimating their baseline results. This does not mean the authors did not consider polynomial choice at all. In fact, in most papers the authors varied the polynomial order as a robustness exercise.

- *Varied BW?* – 'Yes' if the authors varied the bandwidth (to any degree) as part of sensitivity analysis.

- *Varied polynomial?* – 'Yes' if the authors varied the polynomial order (to an degree) as part of sensitivity analysis.

- *Notes* – Additional observations specific to each study.

Table D1: Summaries of applied discontinuity studies published in 2019

Topic	Main outcome(s)	Treatment	Running variable	RD main specification?	Main model	Main function	How bandwidth?	How polynomial?	Varied BW?	Varied polynomial	Notes
[1] Marx, B., Stoker, T. M. & Suri, T. (2019). There Is No Free House: Ethnic Patronage in a Kenyan Slum. *American Economic Journal: Applied Economics, 11:36–70.*											
Ethnic patronage and the rental market in Nairobi	Rent or luminosity of housing	Ethnicity of chief matches own	Distance to administrative boundary	No	Sharp RD	Local linear	CCT	Not discussed	Yes	Yes	RD is used as a robustness check to support claims of exogeneity of treatment variable.
[2] Denning, J. T., Marx, B. M. & Turner, L. J. (2019). ProPelled: The Effects of Grants on Graduation, Earnings, and Welfare. *American Economic Journal: Applied Economics, 11:193–224.*											
College student grants on later economic outcomes	Graduation, earnings, employment	Grant aid	Family income	Yes	Fuzzy RD	Local linear	IK	Not discussed	Yes	Yes	BW not precisely IK – approximately median IK across different years they consider.
[3] Tuttle, C. (2019). Snapping Back: Food Stamp Bans and Criminal Recidivism. *American Economic Journal: Economic Policy, 11:301–327.*											
Ban on SNAP for drug traffickers and recidivism	Recidivism	Drug trafficking offence after cut-off date	Date of offence	Yes	Sharp RD	Local linear	IK	GI argument	Yes	Yes	Bandwidths considered included half IK, CCT and Ludwig-Miller (CV).
[4] Knight, B. & Schiff, N. (2019). The Out-of-State Tuition Distortion. *American Economic Journal: Economic Policy, 11:317–350.*											
Enrollment in state Uni	Enrollment	Being in state	Distance to state border (in bins)	Yes	Sharp spatial RD	Global linear	Not discussed	Not discussed	Yes	No	Theoretical paper with small empirical section.
[5] Dube, A., Giuliano, L. & Leonard, J. (2019). Fairness and Frictions: The Impact of Unequal Raises on Quit Behavior. *American Economic Review, 109:3620–663.*											
Job separation after wage changes	Leave job	Wage step (for Fuzzy RD, average wage of peers)	Distance to wage step	Yes	Cohort-IV and Fuzzy Cohort-IV	Global linear	Not discussed	Not discussed	Yes	Yes	Also do more traditional RD as 'stacked' regressions for each threshold.
[6] Finkelstein, A., Hendren, N. & Shepard, M. (2019). Subsidizing Health Insurance for Low-Income Adults: Evidence from Massachusetts. *American Economic Review, 109:1530–1567.*											
Subsidies and demand for health insurance	Insurance purchase	Poverty thresholds	Income as % of poverty line	Yes	Sharp RD	Local linear	Not discussed	Not discussed	Yes	Yes	Run separate regressions for different income thresholds where subsidy change., Generally use whole range for BW.

[7] Zimmerman, S. D. (2019). Elite Colleges and Upward Mobility to Top Jobs and Top Incomes. *American Economic Review, 109:1–47.*

Elite business schools and labour market outcomes in Chile	Various labour market success measures	Admission to elite business-focussed program	Admission score	Yes	Sharp RD	Mean comparison	CCT	Cross validation	Yes	Yes	

[8] Giuntella, O. & Mazzonna, F. (2019). Sunset Time and the Economic Effects of Social Jetlag: Evidence from US Time Zone Borders. *Journal of Health Economics, 65:210–226.*

Time zones on sleep, and sleep patterns on health and wellbeing	Sleep, health, income	Living across time zone (one hour extra daylight)	Distance to time zone border	Yes	Sharp spatial RD	Local linear	CCT	Not discussed	Yes	Yes	

[9] Nielsen, N. F. (2019). Sick of Retirement?. *Journal of Health Economics, 65:133–152.*

Retirement on health and health care use	Various health and utilisation	Early age pension age and old age pension age	Age	Yes	Fuzzy RD	Local linear	Not discussed	Not discussed	Yes	Yes	

[10] Daysal, N. M., Trandafir, M., & van Ewijk, R. (2019). Low-Risk Isn't No-Risk: Perinatal Treatments and the Health of Low-Income Newborns. *Journal of Health Economics, 64:55–67.*

Birth setting and carer on child mortality	7 and 28 day mortality	OB/GYN attendant and hospital setting (gestational age ¡37 weeks)	Gestational age	Yes	Sharp RD	Local linear	ROT	Visual	Yes	Yes	Guided by ROT BW selection but actually used smaller BW.

[11] Kim, H. B., Lee, S. A. & Lim, W. (2019). Knowing Is Not Half the Battle: Impacts of Information from the National Health Screening Program in Korea. *Journal of Health Economics, 65:1–14.*

Screening threshold for health risks and health behaviours	Healthy behaviours	Cut-offs for risk level categories	Fasting blood sugar, BMI, and LDL cholesterol	Yes	Sharp RD	Local linear	Not discussed	Not discussed	Yes	Yes	The CCT BW is often larger than feasible in their application because it crosses other threshold

[12] Hong, K. Dragan, K. & Glied, S. (2019). Seeing and Hearing: The Impacts of New York City's Universal Pre-kindergarten Program on the Health of Low-Income Children. *Journal of Health Economics, 64:93–107.*

Screening in pre-K for health problems	Health diagnosis various	Eligible for universal pre-K program	Birthdate	Yes	Diff-in-diff sharp RD	Local linear	Not discussed	Not discussed	Yes	Yes	Model is the difference between different RDs. Used IK/CCT in robustness.

[13] David, G., Smith-McLallen, A. & Ukert, B. (2019). The Effect of Predictive Analytics-Driven Interventions on Healthcare Utilization. *Journal of Health Economics, 64:68–79.*

Health risk threshold intervention and health care use	Health care use (ED, hospital, cardiologist, PCP)	Health risk (proprietary algorithm) above cut-off	Given health advice	Yes	Sharp RD	Local quadratic	CCT & IK	Not discussed	Yes	Yes	

[14] Page, L. C., Kehoe, S. S., Castleman, B. L. & Sahadewo, G. A. (2019). More Than Dollars for Scholars: The Impact of the Dell Scholars Program on College Access, Persistence, and Degree Attainment. *Journal of Human Resources, 54:683–725.*

Dell scholars program and success	College enrollment, persistence and graduation	Dell scholars program	Weighted score of achievement, disadvantage and responsibility	Yes	Sharp RD	Local linear	CCT	Visual, significance tests	Yes	No	Paper also uses a DD strategy, which gets equal weight.

[15] Goodman, J., Melkers, J. & Pallais, A. (2019). Can Online Delivery Increase Access to Education?. *Journal of Labor Economics, 37:1–34.*

Enrollment into computer science degree	Course enrollment	GPA above cut-off	GPA	Yes	Fuzzy RD	Local linear	CCT & IK	Not discussed	Yes	No	Uses range of IK and CCT BW to motivate main specification.

[16] Raphael, S. & Rozo, S. V. (2019). Racial Disparities in the Acquisition of Juvenile Arrest Records. *Journal of Labor Economics, 37:S123–S159.*

Police booking on subsequent arrests	Arrests	Age 18	Date arrest relative to 18th birthday	Yes	Fuzzy RD	Local quadratic	Not discussed	Not discussed	Yes	No	RD described is the causal analysis of bookings on subsequent arrest, but sharp RD also used elsewhere for descriptive analysis.

[17] Kreisman, D. & Steinberg, M. P. (2019). The Effect of Increased Funding on Student Achievement: Evidence from Texas's Small District Adjustment. *Journal of Public Economics, 176:118–141.*

School funding and student outcomes	Student achievement (test scores, graduation)	School funding thresholds	District size and sparsity	Yes	Cohort-IV	Global quadratic	Not discussed	Not discussed	No	No	

[18] Frey, A. (2019). Cash Transfers, Clientelism, and Political Enfranchisement: Evidence from Brazil. *Journal of Public Economics, 176:1–17.*

Conditional cash transfer program and political behaviours	Incumbency re-election, various political behaviours	Conditional cash transfer participation	Human development index and municipal population size	Yes	Multivariate fuzzy RD	Local linear	IK for multivariate	Not discussed	Yes	No	

[19] Gallagher, E. A., Gopalan, R. & Grinstein-Weiss, M. (2019). The Effect of Health Insurance on Home Payment Delinquency: Evidence from ACA Marketplace Subsidies. *Journal of Public Economics, 172:67–83.*

Health insurance subsidies and rent and mortgage delinquency	Rent and mortgage delinquency	Subsidy for health insurance	Income as % federal poverty line	Yes	Sharp RD and Fuzzy RD	Local linear and quadratic	CCT	AIC	Yes	Yes	Sharp RD on effect on insurance take-up, and fuzzy on effect of insurance on delinquency.

[20] Farre, L. & Gonzalez, L. (2019). Does Paternity Leave Reduce Fertility?. *Journal of Public Economics, 172:52–66.*

Paid paternity leave and birth spacing	Time between children	Eligibility for PPL	Child's date of birth	Yes	Sharp RD	Local quadratic	Not discussed	Not discussed	Yes	Yes	

[21] Le Barbanchon, T., Rathelot, R. & Roulet, A. (2019). Unemployment Insurance and Reservation Wages: Evidence from Administrative Data. *Journal of Public Economics, 171:1–17.*

Potential UI benefit duration and reservation wages	Reservation wage	Longer potential benefit duration if age 50	Age	No	Fuzzy RD	Not clear	CCT	Not discussed	No	No	

[22] Remmerswaal, M., Boone, J., Bijlsma, M. & Douven, R. (2019). Cost-Sharing Design Matters: A Comparison of the Rebate and Deductible in Healthcare. *Journal of Public Economics, 170:83–97.*

Effect of cost sharing scheme on health care use	Healthcare expenditure	Whether have cost sharing (age 18)	Age	Yes	Diff-in-diff sharp RD	Local linear	Not discussed	Not discussed	Yes	Yes	Estimator is essentially the difference between two RDs.

[23] Scott-Clayton, J. & Zafar, B. (2019). Financial Aid, Debt Management, and Socioeconomic Outcomes: Post-college Effects of Merit-Based Aid. *Journal of Public Economics, 170:68–82.*

Financial aid and later life outcomes	Various socio-economic outcomes: graduation, home ownership, delinquency, neighbourhood, financial security	Received financial aid scholarship	ACT score	Yes	Fuzzy RD	Local linear	Not discussed	Not discussed	Yes	Yes	

[24] Baltrunaite, A., Casarico, A., Profeta, P. & Savio, G. (2019). Let the Voters Choose Women. *Journal of Public Economics, 180.*

Preference voting rules and election outcomes	Female candidate elected	Preference voting and gender quota (municipal pop size > 5000)	Population size	Yes	Sharp RD	Local linear	CCT	Not discussed	Yes	Yes	

[25] Corbi, R., Papaioannou, E. & Surico, P. (2019). Regional Transfer Multipliers. *Review of Economic Studies, 86:1901–1934.*

Local government funding and labour markets	Employment, wages	Municipal government funding	Population size	Yes	Fuzzy RD	Local linear	Not discussed	Not discussed	Yes	Yes	Uses seven different thresholds.

[26] Koster, H. R. A. & van Ommeren, J. (2019). Place-Based Policies and the Housing Market. *Review of Economics and Statistics, 101:400–414.*

Funding to revitalize public housing on house prices	House prices	Eligibility for additional funding for housing redevelopment	Deprivation score	Yes	Fuzzy RD	Local linear	IK	Not discussed	Yes	No	