Dear Professor Sweetman,

Thank you for granting us the opportunity to resubmit the paper and for your extensive comments; we believe they have led us to substantially improve the paper. We respond to your concerns below.

*1.      My (mis-)understanding of earlier versions of the paper was that the data did not allow you to observe individuals longitudinally. However, section 5.3 suggests to me that you can do so. If I am understanding the data correctly now, this has really important implications!*

*a.      You can include lagged demerit points as a regressor.*

Demerit points lagged one period would not be very useful since, in nearly 100% of cases, the point total for the previous day will be equal to that of the current day; moreover, this would assume a linear effect in point balances. We prefer our approach of having current point total categories as being sufficient for purposes of tracking behaviour by the point total up to the previous day.

*b.      You can include driver fixed-effects – though you may not want to.*

We did conduct this analysis as a measure of sensitivity, along with the cluster-robust standard errors discussed below. We did so knowing that this analysis is problematic for a number of reasons. First, the fixed effects annihilate our primary variables of interest: the age and gender categorical variables. Moreover, the policy effects are truly heterogeneous by age group. If we were to have age as a variable (rather than having age categories), the identifying variation would be aging in general, which is not interesting from a policy perspective. For that reason, we redefined the age categories to represent the age of each driver at the time of the policy change to reduce computation time for this sensitivity analysis. This carried with it some problems, since the ageing of drivers is correlated with the policy effect: put simply, all drivers were on average two years older when the policy was in effect. Older drivers get fewer tickets, which introduces a bias to the estimates of the policy effect because it is picking up the effect of aging, which is one of the most important variables in the model.

After including fixed effects, the remaining variation arises from the demerit point balances that change over time during the sample. This causes another problem in the estimation of the fixed effects model because the point balance is essentially a moving average of the points earned over the sample, which is closely related to the dependent variable. We ran some simulations to test this specification and found a that the inclusion of fixed effects causes substantial bias in the coefficients: essentially, most drivers with high positive deviations from their individual average of demerit points have obtained most of their tickets before the policy change, and vice versa for negative deviations, which introduces a negative bias in the policy effect. This problem was apparent in our results from the fixed effects model as the coefficients were between 10 and 100 times as large as those from the other regression specifications. Incidentally, aside from the magnitude of the coefficients, the pattern of effects was sensible: the policy effect was negative and increasing in demerit point balances. However, this effect was noted for both males and females, which is contrary to our unbiased specification, indicating that this bias dominates the true policy effect in the fixed effects regression. In our regressions, we do approximate the individual fixed effect by including the demerit point balance from past tickets for each driver and we feel that this is adequate to condition on a driver’s type, which does change over time as the driver ages.

*c.      Perhaps most importantly, you can adjust the standard errors and statistical tests so that they better reflect the “true” underlying uncertainty regarding the parameter estimates. At present my understanding is that you treat each driver-day as an independent observation, whereas clearly the “true” model has standard errors that are clustered on the driver since those observations are not independent.*

We investigated the sensitivity of the estimates to this richer specification of the error structure. Along with the fixed effects model discussed above, we calculated cluster-robust standard errors, clustered on the individual driver. As expected, the standard errors were slightly larger than those from the standard fixed effects model. The step from the model without fixed effects, however, drives a much larger reduction in the standard errors, owing to the introduction of millions of driver-specific fixed effects. The net result is an estimator with *smaller* standard errors than those in the previous and current version of the manuscript. Considering the substantial bias we found with the fixed effects estimator, we decided not to include them in the paper. We stand by our original specification, with the confidence that our hetero-robust standard errors are not materially misstated, and the estimators are unbiased. We added a paragraph in the section entitled Concerns of Validity to discuss this analysis.

*You should think carefully about (a) and (b) since if I am correct the analysis can be much richer than it is to present. I will focus on the standard errors.*  
*Focusing on the OLS standard errors first. If you can identify all of the driving-days for each driver, then each driver-specific set of driver-days should be clustered together in generating the standard errors. The number of degrees of freedom in statistical test should equal the number of drivers (less k, although in a sample this large that is not particularly important), not the number of driver-days, because of this clustering.*

This point is covered in the analysis that we conducted in response to points (b) and (c) above. Again, we found the standard errors to be *smaller* than before, since the addition of driver-specific intercepts made the model fit the data much better. As mentioned above, we left this out of the paper but mention this sensitivity analysis in the section Concerns of Validity.

It is worth pointing out that these analyses required a considerable effort to program the scripts for the estimation. We generated smaller and simpler datasets to confirm that our results match those produced with Stata. This software has limitations, however, since Stata does not permit different weights on observations within the drivers. In our analysis, however, the data compression method relies on storing the data as frequency-weighted observations, weighted on the driver day: no two drivers with tickets have the same sequence of tickets, point balances and number of days between tickets. Of the drivers with no tickets, the weights could be aggregated more cleanly, but this would also carry with it some challenges because the population of drivers is not fixed: there exist different numbers of drivers without tickets, separated by sex and age groups, on each day in the sample.

*Turning to the logit model, clustered standard errors should also be employed. In the previous version of the paper I assumed that heteroscedasticity-robust standard errors were used for the linear probability model because the best information you had did not allow for the clustering that should have been employed in this context. Further, I did not previously view it as incongruous that homoskedastic standard errors were used for the logit model – despite interpreting what you are doing as quasi-maximum likelihood – since Bernoulli variables are special in the context of quasi-maximum likelihood in not benefiting from sandwich-type standard errors as might other quasi-maximum likelihood estimators such as OLS or the poisson model. (Note that I assume quasi-maximum likelihood, even though neither you nor the referee mentioned this term, since I do not believe the logit distribution is the “true” distribution.) However, even though Bernoulli distributed dependent variables in a quasi-maximum likelihood frame do not benefit from the sandwich-type heteroscedasticity consistent standard error estimator, they do benefit from clustered standard error estimators. Similar to the OLS model statistical tests should be based on degrees of freedom deriving from the number of drivers, not the number of driver-days*

Since we found this course of action not to be fruitful for the linear model, we hesitate to spend time doing the same for the logistic model, not only to avoid the extra effort involved but also because this did not prove to be a worthwhile course of action for the linear model.

*(Aside: Admittedly, statisticians would probably have required a “design effect” adjustment based on some assumed intra-cluster correlation taken from the literature to bump up the standard errors, which are undoubtedly too small previously, but I did advocate for this given the extremely small P-values observed.)*

This all sounds fine, and if we were to rewrite the paper beginning with this approach instead of the way we began two years ago, we might have done so, however, to incorporate any individual-specific covariance structure would require that we rebuild the dataset from raw data. Furthermore, this would preclude much of the data compression that we achieve by aggregating the data and computing regression results using frequency-weighted observations. To incorporate both the time effects and individual effects without aggregation would increase the size of the dataset by a factor of one thousand and would introduce a high degree of computational complexity.

It is worth pointing out that measuring extremely small p-values is not problematic in itself: it is a natural consequence of the size of the dataset and the magnitude of the effect under a consistent statistical test. If the effect is truly not zero, there is no limit on the magnitude of the t-statistics nor the p-values. The size of these p-values, however, does imply that the t-statistics are so large that unless an adjustment increased the standard errors by an order of magnitude, which is unlikely given our analysis mentioned above, the qualitative results would remain unchanged.

*2.      Introducing the MER highlights the importance of nonlinearity and a new dimension of heterogeneity of response to the Québec policy change. I like it a lot.   
a.      I do not entirely understand how/why you defined the MER. It seems to be a marginal effect exclusively for males, as stated on line 352. If this is correct, then the MER in, for example, table 3 needs to be interpreted very carefully. It is a large negative number because it reflects only men, whereas the rest of the table reflects an average of men and women. This is not at all discussed in the text, …*

We chose a male in the MER for the pooled regression because they represent the vast majority of problematic drivers: recall from the paper’s statistics that females are much less likely to be high-point drivers. To calculate the MER, we chose to set the variable to either male or female and did not pick an intermediate value, which would not represent any driver that appears in the dataset. In response to a comment in Point 3 below, however, we removed the pooled results since the model is clearly misspecified and the interpretation of the MER is problematic.

*b.      Moreover, the magnitude of the non-linearity of the patterns of results for the logit regression as seen in the MER suggests – at least in my view – that the linear probability model is probably misspecified (maybe the logit is too since only a very specific form on interaction is automatic to the logit distribution). It seems that it should have interactions between demerit points and age within each sex. Indeed, if you are fan of Angrist and Pischke (as in their Mostly Harmless Econometrics text and other writings) then you would probably want to “fully saturate” the model (at least in terms of age, points and gender) and do predictions and marginal effects based on a fully saturated model. …*

Our analysis generates estimates that are implicitly saturated in that we look at some very specific subsamples (e.g. people ages 23 to 34, the vast majority of our analysis is within sex, etc.), although we do not completely saturate the models. We are aware that the model is misspecified, as all models are, but Angrist and Pischke also argue that models that are not necessarily fully saturated are still best linear approximations. We understand that this methodology would appeal to a small population of readers, however, we prefer a simpler approach to the analysis, which is adequate to address our research question.

We consider the point about the LPM versus the logit model in our response to point 4 below.  
  
*c.      FURTHER:  IN this context, it is not clear how important figures 3, 4 and 5 are. They are taking slices that the data in models that are clearly meant to average across groups ignoring heterogeneity that is known to exist but is different than the issue being focused on in that table. I am not opposed to such tables since policymakers sometimes want to know gross averages, but in the context of the paper each clearly comes from a misspecified model. In some ways, it would be better to present such results as predictions/marginal effects from a correctly specified model.*

The latest submitted draft of the paper does not contain a Figure 4 nor a Figure 5 and you may have meant to refer to Tables 4 and 5. In our response, we justify our choices over both the tables and figures.

We feel that Figure 4 of our latest draft is quite useful to visually capture how the effects of the policy change over time after its initial implementation. This issue was raised by a reviewer in a previous round to understand better the argument surrounding clemency, learning, enforcement, and other issues expected to vary over time.

If, by listing three figures, you are also referring to Figures 1 and 2 in the paper, we could create those using parameters from the full model, rather than simple histograms. The sequence of analysis in our paper progresses from very basic analysis to a richer model and many researchers appear to prefer this storyline. We do see the merit of starting with a single, rich model and then showing all results but, with this approach, readers would be left with the question: What happens if you inspect the data without the model you have chosen? The first two Figures and Table 2, with simple counts of tickets, address this question.

As for the tables, we originally included Table 3 to include pooled regression results, as policymakers may be interested in the overall effectiveness of the legislation. We also included Table 4 to pool the results by age group. We agree that it is not necessary to show the results of these misspecified models and we proceed directly to Table 5, now called Table 3, which presents the results from the separate estimates by gender. These other tables are now in the appendix.

*d.      \*\*Even more importantly, it is not clear to me exactly what model is being estimated in tables 6 or 7. For clarity, it would be good to write down this equation explicitly*

We have clarified the models being estimated in these tables within the text. Those tables show the results from the regression model with the dependent variable indicating that a driver got a ticket with a particular point value. The explanatory variables in this regression do not include the age interactions, which we found were not significant for all point values, since the higher the point value, the greater the proportion of younger drivers. In addition, the values in each row show only the coefficient for the policy effect in the model that predicts the probability of getting a ticket with a particular point value. The equation is now shown in the text, which clarifies the specification.

*3. a)      Somewhat related (1), I do not think you have sufficiently revise the text of the paper in light of the new results. Most glaringly, the introduction and abstract do not put much emphasis on the extremely large effects in the tail of the demerit point distribution, which, again, I think is one of the more interesting findings in the paper.*

We have made an extra effort to fully discuss the new results, including the discussion about the MER and its implications. We added the following sentences to the end of the abstract:

“We find that the new law was effective in deterring motorists from speeding. Moreover, the effect was most pronounced for males compared to females, for young compared to old, and especially so for drivers with high demerit point balances accumulated from past infractions compared to those with few or no tickets. In sum, the change in behaviour was most apparent for those drivers who were the intended targets for the legislation.”

We also revised a paragraph in the introduction as follows:

“We find that the daily probability of receiving a ticket (extensive margin) decreases after the implementation of the law. We first focus our attention on the impact of the policy for different age groups and genders. Young drivers between the ages of 16 and 24 are the most affected by the law,

while there is little effect for drivers over the age of 45. Even though both males and females change their behaviour, the magnitude of the effect on males is about eight times the effect on females. We then consider the effect of the policy on drivers who have a history of speeding. First, we assess the impact of the policy on drivers who had 6 to 10 points on a given day. The effect of the policy is five times larger for this group relative to the average driver. Second, we focus our attention on drivers who have had between 6 and 10 in the pre-sample period and find similar results. Finally, we show that male drivers with more demerit points react more strongly to the policy.”

In the next paragraph, in which we discuss the effects on the intensive margin, we added the following.

“In other words, not only has the policy reduced the number of drivers driving well above the speed limit, it has also led to a decrease in the propensity to commit other moving violations. More importantly, although the effect is noticed for average drivers, who are mainly not speeding, we observe a substantial response from the drivers who are generally more likely to get tickets – those who are appropriate targets for the legislation.”

*3. b) A number of elements of the current paper could be eliminated or moved to an online appendix. Given the evidence in favour of additional interactions, if such a model were estimated, it is no longer clear what the lower panel of table 3 or table 4 contribute.*

Table 3 in the previous draft was our main pooled regression, with all age groups and both genders included. The lower part of the table showed that the effect of the policy varies by age. Our initial thinking was that it would be illustrative to show the pooled results first and then demonstrate how the results are driven by differential behaviour by gender. Ultimately, we decided to eliminate it and split the sample by gender immediately to skip this preliminary analysis of results from a misspecified model.

In Table 4 of the previous draft, we simply re-estimated the model not with age interactions, as we did in Table 3, but with a separate regression for each age group. This was following the suggestion in the last round of revisions in an attempt to clear up any confusion over the apparent differences in results between the logit model and the LPM in the pooled regressions. We agree that this section does not fit well in the paper and decided to move directly to the main results.

*3. c) Even in the current version of the paper, it is not clear what the discussion of cross differences beneath table 4 contributes given the immediately preceding equation illustrating what you do in the paper.*

We enthusiastically agree that this digression interrupts the flow of the paper. We added this discussion only to alleviate the concerns of one of the reviewers who questioned the validity of the regression results after observing differences in interaction terms between the logistic and linear regression models. We have removed this section of the paper and describe only the calculation of the marginal effects that we have presented in the tables.

*4.      Regarding your discussion of the logit versus OLS. My understanding of the traditional debate is that some researchers view the logit as preferable since it provides sensible predictions (not outside the 0/1 balance of a probability), whereas other researchers argue that such predictions are irrelevant in research focusing on marginal effects. Further, those who support OLS suggests that it is more robust since the logit is inconsistent if the logistic functional form is incorrect which rules out heteroscedasticity that many researchers see as ubiquitous (although the quasi-maximum likelihood approach addresses some of these concerns), or if there are omitted variables (regardless of whether or not they are correlated with the included variables), etc. It is not clear to me how strong the evidence is on either side of this divide, but I think there are credible arguments on both sides.*

Our initial submission included only the LPM because the vast majority of papers we have seen with a binary response for the dependent variable used only the LPM and we were just trying to follow convention. We eventually elected to include both the logit and LPM models in the paper because we suspect some readers have strong preferences for one or the other and, indeed, one of the referees was such a reader with strong preferences for the logistic regression. Angrist seems to be quite in favour of always running LPM, and he once said “[OLS] always predicts y as well as possible, and y can be Bernoulli or non-negative or whatever, … I think regression is always appropriate”.

The most compelling reason to show results from both models, regardless of the researcher’s preference, is that the policy effect is measured both as a proportional change in probability, as with the logistic model, and as a fixed change in percentage points, as with the linear probability model. This is important in our context because of the large gender differences in the baseline probability of getting a ticket; otherwise, the reader would be left with the question: There exists a difference in terms of percentage points, but is the difference still apparent when measured as a proportional change in probability? We did not adequately emphasize this in the earlier version, but we now highlight this in the opening sentences of the Empirical Results section.

*5.      In places the writing could be improved. I am not talking about the grammar/spelling, but what is said. Please read the manuscript carefully. Esp. the abstract/intro does not seem to emphasize the most interesting new findings - the intensive/extensive discussion seems like an effort on this dimension but seems too opaque to me. Here are a few examples of specific types of other things you should address.*

We made our best effort to modify parts of the paper to address the points you made. In the previous submitted draft, we focused our attention on the methods section and neglected to add some language in the abstract and introduction to reflect the new findings. We apologize profusely for this oversight.

*a. on line 359 it says “very precisely estimated (t-statistic equal -94.054)”. Should you not be referring to a standard error instead of a T-statistic? The t-statistic might be large because the truth is far away from the hypothesized value regardless of whether the coefficient is precisely estimated or not, similarly the t-statistic might be very small if the truth is very close to the hypothesized value regardless of whether the standard error is large or small.*

We cut the offending sentence.

*At line 364: “the effect loses its statistical significance”, but the coefficient being discussed measures the difference between the relevant age group and the under 16 age group. The total effect for the relevant age group is the sum of the two coefficients (POLICY AND POLICY\*AGEGRP). It is not clear whether the total effect (i.e. the sum of the two coefficients) is statistically different from zero or not.*

We initially changed this sentence to state that “the *age difference* (emphasis added) in the policy effect is not statistically significant.” In response to comments raised above, however, we have moved this section to an appendix.

*b. i. Having the omitted group as under age 16 drivers seems odd. How many under age 16 licensed drivers are there in Québec? This seems like a very special group. Do they have standard driver’s licenses?*

Yes, these drivers can have standard driver’s licenses in the time period under examination. We chose this group because they seemed like a natural “neutral” group that does not drive very much (they aren’t even old enough to be out of high school or hold a full-time job) and thus have a much smaller chance of being affected by the policy as a result. In our view, this group serves as a sensible benchmark because these drivers have a limited driving history.

*b. ii.     Relatedly, keep in mind that in, for example, the bottom half of table 3 the coefficient on the “policy” variable is measuring the effect of the policy on those under age 16.*

Yes, this is correct, however, these coefficients are not statistically significant. We modified the descriptions of the tables to clearly state that the ag-policy interaction coefficients represent the difference from the benchmark age groups.

*c.      At line 367 it says “representative drivers”, which is true but it would be good to remind readers that the representative driver is a male – this seems to get lost to me as a reader at times.*

We added additional context in the paper to help avoid confusion. We did add clarifying statements for the remaining tables, however, the potential for confusion is mitigated by the separate estimation for males and females.

*Finally, let me say once again that I do really like this paper. I think the increased focus on what I will term the “right tail” of the demerit point distribution is a very interesting and important addition to the most recent manuscript revision. This is what I take to be the key point: Even though the average driver might not be much affected by the policy change (which makes sense since the average driver is not speeding excessively), those who are appropriate targets for the legislation have very substantial policy/behavioural responses. I want to publish the paper since I think it speaks to an important Québec policy issue and I like the idea of advocating for research that focuses not on the average effect of a policy change, but on the effect for the targeted population for the policy intervention. However, I am not yet convinced that the analysis meets the standards of the CJE. It may be that I am misunderstanding some elements of the paper, but I have put in appreciable effort to understand the paper and if I have misunderstandings then other readers are likely to have them as well. Also, if I am correct about (1), then you should think about (1a) and (1b), and, in particular, some improvement inference (1c) would be worthwhile.*

We agree on the importance to focus on this “right tail” of the demerit point distribution; please see our response to 3.a. an example of an insertion we made concerning this.

We thank you again for your comments and for allowing us to revise our work.

Best,

Vincent Chandler

Lealand Morin

Jeffrey Penney