# Response Plan

Comments in light blue are suggestions for the next version of this document.

## Point 1.

*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!*

We lack the computational capabilities to implement many of the standard panel models as the full data set would take up hundreds of gigabytes of memory, and so computation time and even capability (e.g., some matrices may be too difficult to invert, issues with numeric overflow, etc.) prove genuine obstacles. In fact, we did not even create the full dataset and only know its approximate size since the aggregation of the dataset into frequency-weighted observations reduces the number of rows by a factor of a thousand. Even with the weighted dataset, computational issues still prove to be problematic.

*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 previous period’s point total will be equal to that of the current period; 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 current point total.

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

The fixed effects would absorb our primary variables of interest: the age and gender categorical variables. Moreover, the effects are heterogeneous by age group. If we were to have age as a variable (rather than having age categories), the identifying variation would be ageing in general, which is not interesting from a policy perspective.

*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.*

The software and computing capabilities to calculate the standard errors on data sets of this magnitude might not and probably do not exist; the long data set cannot even be loaded into Stata. We feel that hetero-robust standard errors are sufficient. For experts in the field of big data, many are not even familiar with the concepts of hetero-robust standard errors, and may not have even heard those terms spoken before. The reason for the apparent gap in the computational literature is that, when the dataset is large, model specification is generally seen as much more important than the standard errors. In other words, the first moment (the numerator of the t-statistic) is more important than the second moment (the denominator), for which accuracy is irrelevant if the specification is not correct.

## Point 2.

*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. We will be much clearer in the text as to what the MER is and means for each table in the text.

*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 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, but Angrist and Pischke also say that models that are not necessarily fully saturated are still best linear approximations. We can modify our analysis to add more interactions, but we are unsure as to what they would add to the analysis.

We consider the point of the LPM versus logit below in point 4.  
  
*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 or a Figure 5.

We feel that Figure 3 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 better understand 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 might 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 are left with the question: What happens if you inspect the data without the model you have chosen?

*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 will clarify the models being estimated on these tables in the text.

## Point 3.

*3.      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, …*

*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.    
  
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.*

Table 3 is our main pooled regression, with all age groups and both genders included; the lower part of the table shows that the effect of the policy varies by age. We could in theory eliminate it and split the sample by gender immediately to save space, but our initial thinking was that we would be illustrative to show the pooled results first and then demonstrate how the results are driven by differential behaviour by gender.

In Table 4, we simply re-estimated the model without age interactions, as in Table 3, but ran 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 will make an extra effort to fully discuss the new results, including the discussion about MER.

## Point 4.

*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 where the dependent variable was a binary response 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. 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, with the logistic model, and as a fixed change in percentage points, 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 highlight this in the paper, but it is worth mentioning in more detail.

## Point 5.

*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 will work hard to rewrite substantial 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/introduction to reflect new findings. We apologize profusely.

*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?*

Note: I am specifically addressing this point because it requires a specific response; the other points under (5) point to specific issues we can easily fix.

Yes, they 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 has a much smaller chance of being affected by the policy as a result.