Response to "Income, the Earned Income Tax Credit and Infant Health: A Comment", Submitted to *AEJ Policy* by Daniel Dench and Ted Joyce

By Hilary Hoynes, Douglas Miller and David Simon
August 10, 2018

Thank you for the opportunity to respond to the comment by Dench and Joyce (hereafter, DJ) on our 2015 AEJ Policy paper (hereafter, HMS).

We summarize the comment as making 3 main points:

1. Pre-trends for OBRA 1993 difference-in-difference and event study analysis
2. Sensitivity to adding quadratic trends by parity to panel fixed effects model using multiple expansions
3. Appropriate clustering of standard errors

Below we address these three main comments followed by our response to other comments.

Before going on to discuss these three points, we wanted to point out one change in our results presented here compared to our published paper. In reengaging with the HMS code we noticed that we had failed to control for the full set of demographic cells in our model (HMS equation 1, p. 183 "we include fixed effects for demographic group ϕ_j"); in particular we had missed some of the third order interactions. We corrected this error and all results in this response use the correct specification. Not surprisingly (given they are fixed effects) in practice this makes little differences in our estimates. For example, in HMS Table 2 column 1 adding the full set of demographic fixed effects changes the estimate to -0.359 from -0.355.

<u>Response to DJ Main Comment 1: Pre-trends for OBRA 1993 difference-in-difference and event study analysis</u>

Our paper uses variation in the federal EITC to estimate the effects of the credit on infant health (low birth weight). We use two identification strategies. Our main approach is to leverage the variation from the 1993 expansion of the federal credit, which varied by parity (of birth) and tax year. We use difference-in-difference and event study models – in both models we form treatment and control groups using parity of birth (e.g. compare parity 2+ to parity 1).

A central criticism in DJ is that HMS use a pre-period of 1991-1993 to estimate the OBRA 1993 difference-in-differences and event study models.

Why did we choose this time period: (1) Policy Changes -- Our choice of 1991-1993 for the pre-period was a deliberate, pre-determined choice, based on the timing of the tax change(s) and following the practice in the existing literature.

First, as we show in HMS Figure 1, OBRA93 is the largest expansion of the EITC and also the only one where the policy expands differentially across three parity groups (1, 2, 3+ corresponding to childless, one child, and two or more children). Additionally, HMS Figure 1 shows that the OBRA93 expansion took place between 1993 and 1994 for parity 1 and parity 2. For parity 3, the tax change was phased in

between 1994 to 1996. HMS Figure 1 also shows a previous tax change (OBRA90) that expanded the credit for parity 2 or more (childless filers, those of parity 1, have no EITC until 1994). It is because of this previous tax change that we use 1991-1993 as our pre-period for analyzing OBRA93. DJ re-estimate the event study models using 1983-1993 as the pre-period, in part justified by their view that "1983-1998 provide(s) a more complete picture of the time-series patterns in the DD analysis." (DJ p. 6). The first panel of the table below shows the difference (relative to 1993) in the maximum credit for each family size (parity) and the second panel of the table shows the difference-in-difference (relative to 1993, relative to parity 1). We also show the difference-in-difference in Response Figure 1.

Looking at the first panel (difference relative to 1993), the table shows that the gain in the maximum credit for those with 2 or more children rose by $199 between 1991 and 1993. However, between 1990 and 1991 it rose by $300. There is a similar pattern for those with 1 child. It is this large increase between 1990 and 1991 that led us to set the pre-period where we did.

Notably, Response Table 1 and Response Figure 1 also show that the OBRA90 expansions were quite similar for parity 2 and parity 3+ while the 1993 expansion was much larger for parity 3+. This illustrates why HMS present extensions of the "DD3" specification (the comparison of parity 3+ to parity 2) and use a pre-period back to 1987 (HMS Fig 3b).

Why did we choose this time period: (2) EITC Literature -- Additionally, our choice to use OBRA93 -- in a difference-in-difference model and with a (short) pre-period – also follows other papers in this literature. For example, Evans and Garthwaite (2014) use a similar approach to examine impacts on maternal health.

What is new in DJ's critique? The issue of the longer term trends in the period prior to OBRA93 is presented and discussed in the paper. We begin the paper by presenting the difference-in-difference results (HMS Table 2) along with the event study version of these results (HMS Figure 3). After presenting these main DD results, we discuss them within the context of longer term trends:

> "We provide additional context for these results in online Appendix Figure 1 panels A–D. In these figures we show raw trends in the rates of low birth weight by parity, over the period 1981–1999. In each figure we show low birth weight probabilities relative to the 1993 level. Online Appendix Figure 1, panel A has results for our high-impact group, and shows two main findings. First, the changes we observe in our experimental period occur within long-run decreases in rates of low birth weight for this population (although in the pre-1991 period there are many other policies changing, including earlier expansions of the EITC). Second, the raw trends for 1991–1998 show the same pattern as in our main event study results in Figure 3, panels A–C; thus these results are not strongly impacted by the covariate controls in the model. In contrast, the trend for all births is one of increasing rates of low birth weight (online Appendix Figure 1, panel B). The long-run trend for white high-impact mothers (online Appendix Figure 1, panel C) is similar to that for all high-impact mothers, while for high-impact black mothers (online Appendix Figure 1, panel D), the trends are less monotonic. All the figures show that there are large changes in parity gaps in low birth weight that occur during the 1980s. It would be valuable, but outside of the scope of this study, to understand these changes better." [HMS p. 187-188]

In sum, our choice of the pre period is the direct result of the tax law changes and this approach is commonly used in the EITC literature. Additionally, HMS are clear in pointing out how this period fit into

longer term trends. We remain of the view that explaining these longer run trends is outside the scope of our paper. However, in response to DJ, we have taken a deeper dive to see if we have any insights into the trends in low birth weight since the early 1980s.

Longer term trends: what it isn't. First, we can rule out what it isn't. In HMS Table 10 and Figure 6, we explored changes in the fertility and the composition of births. We found that the sample, within our research design of comparisons across parity, is getting older and more educated over time. However, we find these changes are operating smoothly through the 1991-1998 period which rules out a change in selection of births around the timing of OBRA93 expansion. To explore these longer term trends, we have extended that analysis in several ways. First, we estimate a low birth weight prediction model using the full vital statistics sample (1983-1998) with a full set of interacted demographics on marital status (married, unmarried); ethnicity (Hispanic, non-Hispanic, missing); race (white, black, other, and missing); educational attainment (less than high school, high school graduate, some college, 4-year degree or higher, and missing education); and age group (18-24, 24-34, and 35 years and older). We use this to predict low birth weight for each observation in our estimation sample. We then estimate event study models of predicted low birth weight over 1983-1998 for all races and for blacks and whites. These are presented in Response Figure 2 which shows DD1 (parity 2+ vs 1) and DD3 (parity 3+ vs 2). These figures show no consistent pattern of changes in composition of births. The estimates move around quite a bit but do not echo the longer term patterns in low birth weight. This exercise tells us that these observable demographics are not driving the trends in low birthweight during the 1980's and 1990's.

Second, in Response Figure 3, we also present event study graphs for log(births) over 1983-1998 to extend our analysis of changes in fertility (HMS Table 10, column 1). Here, consistent with the results in HMS Table 10 Column 1, we find little change in fertility across parity over this time period. This exercise presents evidence against differential selection-into-fertility driving the trends in low birthweight during the 1980's and 1990's.

Based on this analysis, we rule out any changes in number or composition of births across parities during this time period.

Longer term trends: importance of other policy controls. To explore the factors other than the EITC that might be affecting trends in low birth weight over the longer term, we extend our baseline models in two ways. First, in HMS, we include three state by year varying variables (unemployment rate, Medicaid/SCHIP eligibility, and welfare reform dummy). We contacted several experts to inquire about other possible controls. We settled on using the extensive controls included in Kearney and Levine (2015).[1] Second, we recognized that our original specification (see HMS equation 1, and in particular the single coefficient for each state x year policy variable X_{st}) was not well specified to absorb variation *by parity* (and year), which is of course critical in our design (for example see Meyer and Rosenbaum 2001 and Hoynes and Patel 2018). Here, we allow the state x year varying policy variables to have different impacts across parity (HMS equation 1 is modified to estimate β_j rather than β).

[1] Compared to the full set of policy variables analyzed in Kearney and Levine (2015), we exclude measures of mandatory state HIV/sex education policies (which is only available for 1998) and state expenditures on child support enforcement (which is only available in 1990). We also exclude whether a state receives Title V-510 abstinence education funding, a provision passed as part of 1996 welfare reform. Only two states (CA and NH) accepted this funding by 1998 (and both did so in 1998).

Response Figure 4 shows the event studies for DD1 (2+ parity compared to parity 1) and DD3 (parity 3+ compared to parity 1) for all, blacks and whites. These figures show that adding these controls and allowing their effects to vary across parity leads to some reduction in the pre-trends for comparisons of parity 2+ versus 1. For example, for the high impact sample ("all") adding these controls leads to a fairly flat pre-trend back to 1987. Overall, there is less impact on the comparisons of parity 3+ versus 2. Notably, if anything, the pre-trends for parity 3+ versus 2 on the whole sample show *increases* in the incidence of LBW which if continued would bias our estimates toward zero.

DJ point out that the largest pre-trends were for black mothers, the group where HMS see the largest impacts. The second row of Response Figure 4 shows the extended pre-period for black mothers for DD1 and DD3. For DD3, the pre-trend is flat going back to 1990. Prior to 1990 the trend is **upward**, which if anything makes the post 1993 decline more surprising. For DD1, there are some fluctuations around zero through the pre-period, but none are statistically significant.

Across all the figures in Response Figure 4, while the pre-trends change in direction and amplitude depending on the subsample and model looked at, there is little qualitative change to the post-93 impact, and if anything the post period suggests larger effects than in HMS.

In addition to these models, we also explored two other extensions to our event study to explore the longer term trends. First, we expanded the controls to include demographic cell x parity fixed effects (rather than one way demographic cell and parity fixed effects). Second, to account for possible selection on fertility, we included a control for log(births) in the cell. If there were systematic changes in selection-into-fertility or long term trends in health that were associated with low birth weight, and that also were correlated with the policy expansion, we would expect this control to have an impact on the policy coefficient. The results were unchanged with these additions.

Ultimately, this analysis does not fully explain the long term trends in low birthweight. Our work does demonstrate the sensitivity of the long term trends to policy-by-parity controls and we believe this is the type of analysis needed to shed more light on what was happening in the period DJ highlight as a potential area for concern. Notably, the post-1993 treatment effects are little changed by these controls.

Longer term trends: what are trends in smoking and prenatal care? To get the full picture of changes over this longer time period, we extend the analysis of the variables smoking and prenatal care that we explored as possible mechanisms of the EITC in HMS (HMS Table 7)[2]. We extend that analysis in two ways. First, we estimate event study models over the full period (following the approach above). Second, we add the state x year policy variables and allow their effects to vary by parity (as above). In particular, we estimate event study models of smoking (HMS Table 7 column 6) and any prenatal care before the 3rd trimester (HMS Table 7, column 1). Smoking is available in all states starting in 1987 so our analysis of that outcome variable spans 1987-1998.[3]

Response Figure 5 plots the event study results for smoking. Overall, the evidence is quite consistent with the OBRA93 incentives: the incidence of smoking turns down post-1993. The pre-trends for the

[2] In DJ footnote 7 they extend their critique of the choice of 1991-1993 as the pre-period to our analysis of mechanisms, pre natal care and smoking.

[3] To be more clear, smoking is available in all states and years starting in calendar year 1989, which based on our "effective tax year" means we can go back to 1987.

comparison of parity 2+ versus parity 1 are relatively flat pre-1993 for all, black and white. The comparison of parity 3+ versus parity 2 shows flat pre-trend for blacks back to 1989 and for all and whites back to 1992; there is, however, an unexplained upward trend before that. Interestingly, for whites, smoking is trending up through 1987-1993 for parity 3+ versus parity 2 which could explain why we fail to find an effect of the EITC on whites.

Response Figure 6 plots the event study results for pre-natal care (any care before 3rd trimester). Overall, these figures show confirmatory evidence of a flat pre-period near the 1993 expansion followed by an increase in visits post-1993. There remains a decline in pre-natal care in the early part of this period (1983-1987) that our state-year control variables do not explain.

Summary: In our view, our choice of a transparent design applied to the largest and parity-varying expansion in the EITC (OBRA93) remains the best approach to estimating the effects of the EITC on birth outcomes. We chose the pre-period based on existing policy variation and following the literature. The extension of HMS to control for other policy changes over this time period, explained some but not all of the trends over the longer period. In contrast, the additional controls do not lead to much change in the post-1993 estimates. Overall, more attention to understanding the longer trends are of value and should be analyzed in future work.

We note that including our set of policy-by-parity controls changes the long term pre-trends in the event study. However the short term flat pre-period from 1991-1993, as well as the downward inflection in low birthweight in the 1994-1996 period are more robust to the choice of policy-by-parity controls. We take this as evidence that the potential pattern of "mean reversion" central to DJ's concerns is sensitive to this richer set of controls, and therefore it is unlikely that HMS results are driven by mean reversion. Instead, we view these results as evidence that a number of important policy changes and trends related to fertility and low birthweight were occurring in the 1980s, but that HMS's estimates likely reflects the best estimate of the effect of the OBRA93 expansion.

<u>Response to DJ Main Comment 2: Sensitivity to adding quadratic trends by parity to panel fixed effects model using multiple expansions</u>

While the OBRA93 analysis is at the core of HMS, we also present estimates from a panel fixed effects model to allow for leveraging the variation across multiple EITC expansions. We estimate a parametric model where we control for the maximum credit (varying by parity and year) along with a full set of year and parity fixed effects (as well as demographic cell fixed effects, and state-year policy controls). This model is presented in HMS equation 2.

DJ critique this approach by showing the sensitivity of the results to adding controls for parity x linear trend and parity x quadratic trends (DJ Table 1).

Relationship to DJ Main comment 1 (pre-trends): This critique comes down to the same fundamental point raised in the issue above -- the role of the longer term trends in LBW.

What is new in DJ's critique? The issue of sensitivity to parity trends is presented and discussed in the paper. In HMS Table 5 we show the sensitivity of the results to adding controls for parity by linear time. In discussing these results, we say:

> "Due to the longer time span, with multiple EITC expansions, we can explore the sensitivity of the results to the inclusion of parity-specific linear trend (in year). The results (in columns 2, 4, and 6) show substantially larger estimates treatment effects for model with parity linear trends. While we may be "overfitting" the parity-time relationship, we view the robustness to including the parity trends as an important result." (HMS p. 195)

DJ expand on this approach and also control for parity by quadratic linear time. On page 3 DJ refer to the inclusion of quadratic trends by parity as a "*modest addition to their specification*." While it is true that this only involves three additional parameters, because the majority of the EITC occurs with the OBRA93 expansion and the time series of the pre-period is relatively short (1983-1993), this represents a considerable burden on the data to estimate quadratic trend terms *separately from the effects of the policy*.

Presumably DJ's intention in adding quadratic controls is to control for "pre-existing trends"; trends that would have happened even if the policy had not changed. However, because the policy does change frequently over time, and in broadly non-linear fashion (see Response Figure 1 above), there is a nontrivial risk that the quadratic trends will be over-fit to capture the impact of the policy. This would not happen if the trends were fit only to time periods with no policy impacts. But with a global polynomial fit there is a risk of the trends becoming "bad controls", in the sense of being fit to the causal impacts of the policy changes.

Finally, quadratic trends is not a standard specification in this literature. Given concerns about the short length of the time series, the many periods of change, and the potential for overfitting, it is unclear whether the quadratic should be preferred over linear or no trends.

<u>Response to DJ Main Comment 3: standard errors</u>

DJ express concerns with the approach to statistical inference in HMS.

What is new in DJ's critique? In response to requests from the editor and referees (during the review process of HMS), we explored the sensitivity of our results to the level of clustering. As we note in HMS (p. 183) "We have also explored alternative levels of clustering, including: demographic group level, parity-by-year, two way clustering on demographic group and state, and two way clustering on parity and state. Statistical inference results are robust to these alternative clustering choices." We reproduce here the "response table" associated with these findings in Response Table 2.

DJ propose two alternative approaches: (1) collapsing to pre-post by parity; and (2) an approach adopted from Donald and Lang.

DJ are correct to note that the traditional difference in difference framework relies on assuming independence of errors across pre-post for treated and untreated groups. It is also true that when a researcher has many different cross-section units, that this assumption can be relaxed. One way to relax that assumption would be to "cluster on cross-sectional unit". Another way would be to collapse each cross-section unit's data to pre-post; and then estimate the model on this collapsed data. Both of these approaches are appropriate when there are many cross sectional units.

DJ's proposal to collapse the data to pre-post is only appropriate when there are many cross sectional units. In our case, there are 4. (The 4 cross section units are: parity = 1, parity = 2, parity = 3, and parity

= 4+. And if we consider distinct levels of EITC treatment as a unit, then there are only 3 units.) It is inappropriate to adopt this pre-post collapsing approach with so few cross-section units.

To see this point more clearly, let's suppose that we had two cross-section units: untreated (parity 1) and treated (2nd and higher parity). If we were to collapse the data to pre-post, we would have 4 observations, and a regression with 4 terms (constant, post, treated, and post*treated). We would not think of using this regression to obtain statistical inference. Instead, what would be common practice would be to estimate the model on the micro data; the statistical uncertainty in this model comes from the sampling uncertainty of the averages for each of the 2x2 cells. That approach requires the assumption of independence across the 2x2 cells. This assumption is standard in these type of difference-in-difference models.

Our data setting is not a 2x2 design; instead it is more like a 3 (groups) x 8 (time periods) design. One approach sometimes employed in this setting is to cluster on time. Here we also explore the robustness to clustering on time period, with no important changes to the statistical inference. Results are presented in Response Table 3.

DJ propose a second approach borrowed from Donald and Lang (2007). There are many approaches presented in Donald and Lang, and DJ are not clear about which one they employ. We assume it is the approach outlined in Donald and Lang pg 227-228 in the reanalysis of Gruber and Poterba, based on the text on page 228 "If instead we assume that the differences in the differences between the employment rates is i.i.d., in other words that the differences are a possibly correlated, random walk and normal, we can estimate the regression using differences."

This approach relies on comparing the 1-year D-in-D estimate (1994-1993) to the distribution of other 1-year D-in-D estimates. The idea is that the other D-in-D estimates provide a distribution of what we might see when there is no policy change. So we can see how the 1993-1994 change fits into the distribution of "no policy change" estimates. This is an appropriate exercise only if the other years are truly not impacted by policy changes. Further, it is an approach that throws away much of the variation (it focuses on a 1-year frequency change, throwing out longer-change comparisons). In our context, there are several reasons why this is an inappropriate model. First, the 1993 expansion was phased in over a few years. Second, there might be some delayed behavioral responses to the policy. Third, while we aim to make the most accurate mapping of birth timing to policy, there may be some slight mis-match. All of these reasons mean that the "no policy change" D-in-D comparison years might in fact be tainted by policy change – and so they do not provide a valid counterfactual distribution.

The results in DJ's table 3 are telling – the point estimates are radically different than the main point estimates in HMS. For example DJ table 3, column 1, black mothers, the point estimate is -0.198 (se=0.239). The corresponding estimate from HMS Table 3 is -0.728 (se=.143). So while the Donald and Lang method is presented as a way to adjust inference, what is really going on is that it is a different estimator.

(As an aside: this change in treatment effect estimate stands in contrast to the Gruber-Poterba replication in Donald and Lang. Donald and Lang are able to replicate the main point estimates using this technique. The inability to do so in our application may be further evidence that it is an inappropriate method for our application.)

In the same section in the Donald and Lang paper, they also propose an alternate test, focusing on the Treatment-Control differences for each year. They propose regressing these differences on an indicator for post change. When we use this approach comparing parity 2+ to parity 1 for black children (based on the numbers in DJ Appendix table A-2) we obtain a treatment estimate of -0.757 (se=0.255). This approach (also proposed by Donald and Lang) has the desirable property of recapitulating the main treatment effects reported in the paper (-0.728 (se=0.143)). And while the standard errors are larger than those originally estimated in the paper, they do not lead to substantively different conclusions regarding statistical significance.

To summarize our discussion regarding statistical inference:

- DJ make a correct observation that Difference-in-Difference models with few cross-section units require stronger assumptions of independence than when there are many cross-section units.
- However, this observation is not novel; the standard approach for D-in-D with few cross-section units embodies this assumption.
- We have tried several alternative clustering approaches; our results are robust to all of these approaches.
- DJ's proposal to collapse to pre-post is inappropriate with so few (3 or 4) cross-sectional units.
- DJ's proposal (taken from Donald and Lang) to compare 2-period D-in-D coefficients is poorly suited for this application.
    - The method is only appropriate if treatment is a clean step-function (no ramp-up); and if response-to-treatment is instantaneous (no delayed treatment effects).
    - The method does not only impact measures of statistical uncertainty: instead it produces a much smaller estimated treatment effect. This is what drives the reduction in statistical significance.
- An alternative approach proposed in Donald and Lang recapitulates the main treatment effects, and the impact on the estimated standard errors is insufficient to change any conclusions.

Below we provide responses to the other points raised by DJ (which we read as less central to their main argument):

(The numbering is ours.)

1) [DJ p. 2] "*HMS Aggregate the census of individual birth certificates into cells by state, year, parity, race, age ethnicity and education. However, the EITC varied only by year and parity. We demonstrate that thousands of their cells are superfluous.*"

   We state in several places in the paper what the level of variation is in the paper. For example, on HMS p. 174: "This commonly used approach in the EITC and labor supply literature leverages variation across time and across family size."

   Further, just because the exogenous variation from the policy occurs at the level of year and parity does not mean that all other cells are superfluous. Just as, when running an analysis on individual data, it does not make sense to collapse down to the level of exogenous variation. Including additional cells allows us to control more richly and flexibly for demographic differences, which allows for cleaner identification of trends and greater precision. This also allows us to explore differences across groups with varying levels of propensity to be affected by the policy (see HMS Tables 3-5, Figure 4).

2) [DJ pp. 4-5] "*Lastly, HMS's identifying variation comes from differences in infant health by parity overtime at the national level. The relatively smooth growth in EITC benefits and participants since 1986 makes identifying small effects of the program from underlying trends even more challenging (Figure 1).*" And relatedly [DJ Section III] "*They limit the pre-period to three years to exclude effect from the EITC expansion in 1990 (HMS p.187). However, the 1990 expansion was phased in over three years much like the 1993 expansion. In fact the year-to-year increase in tax filers for the EITC and the average tax credit per family from 1991-1993 are similar to those from 1994 to 1996 (See Figure 1). Not only is 1991-1993 a limited period with which to determine pre-existing trends in low birthweight, but the smooth growth in EITC participants and tax credit refunds attenuated the discontinuity in 1993 and jeopardizes the parallel trend assumption*." And finally, in footnote 5 "*Between 1989 and 1993 the average tax credit per family increased by $464 ($1028-$564) and the number of tax filers grew by 3,421 million. The changes were almost the same between 1993 and 1997. The average tax credit per family increased by $537 ($1567-$1028) and the number tax filers grew by 4,217 million*."

   There are several issues to discuss here:

   "However, the 1990 expansion was phased in over three years much like the 1993 expansion." We discussed this point in our response to main point 1 (OBRA93 pre period). It is clear that the 1990 OBRA expansion largely took place in one year, with a modest phase in.

   "In fact the year-to-year increase in tax filers for the EITC and the average tax credit per family from 1991-1993 are similar to those from 1994 to 1996 … the smooth growth in EITC participants and tax credit refunds attenuated the discontinuity in 1993 and jeopardizes the parallel trend assumption." DJ Figure 1 plots the <u>total number of filers</u> and the <u>average credit received per EITC filing unit</u>. This figure (DJ Figure 1) does not provide the data needed to assess this issue.

First, the average credit is presented in nominal rather than real dollars. Second, the composition of the EITC recipient population changes dramatically in 1994 when the childless credit is introduced. The number of childless filers is sizable, but their benefit is very low; thus as the childless grow as a group, they pull down the average benefit, obscuring the trends for other groups. To use data on average credit per filer to inform this issue, you must report trends separately by family size. We do that in Response Figure 7 -- which shows the trend in number of filers and average real benefit, separately by number of EITC children

Once you analyze the data by number of children, "*the smooth growth in EITC participants and tax credit refunds attenuated the discontinuity in 1993"* is not true. The graph shows that average EITC per filer is similar pre-1994 (consistent with the policy parameters), is nearly identical across parity -- the basis of our design, and that the increase ramps up over several years for the larger families (2+ children).

With respect to interpreting the time series of counts of total EITC tax filers - DJ use this data to argue that the 1990 expansion increased the number of filers as much as 1993 did. This is problematic for several reasons. This is also refuted by breaking it down by family size. The top figure of Response Figure 7 shows that number of filers with one child increases 1990-1993 but is flat post 1993. And in contrast the number of filers with 2 or more children increased throughout this period perhaps more steeply post-1993. This too is consistent with the policy changes. It turns out that the 1993 expansion increased the earnings eligible range for those with 2 or more children but did not expand the earnings eligibility range for those with 1 child (see Response Figure 8). Therefore, for 1 child filers, OBRA93 led to an increase in the average credit amounts but the number of filers did not change substantially. Further, and more importantly, this time series trends are determined by many things in addition to tax policy including population growth, demographic change, business cycle fluctuations (Bitler, Hoynes and Kuka 2017), etc. This is why we leverage policy variation the way we do.

3) [DJ p. 4] DJ states "*Why an income transfer would only affect black women and not whites or Hispanics of the same SES and parity is difficult to explain.*"

First, we do not think that the findings are that only black women are affected. While the point estimates for whites and Hispanics are smaller and the results are not statistically significant, we believe that the confidence intervals include meaningful effects. DJ are perhaps interpreting an insignificant coefficient as a precise zero. Second, the baseline distribution of birthweight is lower for black children. This results in higher rates of low birth weight; and also may lead a similar sized shock to birthweight to result in greater changes to low birth weight. Third, Hispanic mothers may have a lower effective treatment dose from the policy, because a larger fraction of these mothers are undocumented immigrants who do not qualify for the EITC. Further, Hispanic children have low rates of low birth weight to begin with; this should result in a smaller coefficient. (These points were already made in HMS pages 188-189.) Finally, it is plausible that even conditioning on education and marital status, that black women face higher levels of disadvantage than their white counterparts. In short, the patterns of results are not inherently problematic.

4) [DJ p. 5] "*The difference in low birth weight of women of parity 2 versus parity 1 is smaller, 0.164 percentage points. These are small effects of questionable clinical relevance*."

A central point in HMS is that the expansion was larger for those having their third or higher order birth compared to those having their second birth. Thus, we should expect a smaller effect for 2nd births. For example on HMS p. 185 "Since the OBRA93 expansion was larger for families that already had two children, the second model (shown in column 3) decomposes the policy impact into second births and third or more births. The results are consistent with expectations: low birth weight status is reduced by roughly 0.53 percentage points for third and higher order births versus -0.16 for second births (each compared to first births.)" We also show the first stage (on after tax income) in HMS Appendix Table 1 and use this to convert the intent-to-treat estimates in DD model into the treatment-on-the-treated estimates (HMS Table 4).

We do not believe that "clinical relevance" is the appropriate standard for assessing the magnitude of the effects, as this study is not intended to guide any particular clinical decision. Instead, the magnitudes of the effects should be interpreted in the aggregate. We discuss these aggregate magnitudes in terms of economic significance in HMS pp 206-207. In our assessment, the comment does not add to the discussion in this area.

5) [DJ footnote 7, p. 8] "*Finally, smoking needs to be an inferior good among low income women of parity 2+ relative to parity 1 for there to be an inverse relationship between the income tax credit and prenatal smoking.*"

We discuss this point in the paper (see HMS, pp. 177-178). We make the point that behaviors such as smoking and drinking could change due to increases in after tax income (e.g. smoking is a normal good). But we additionally point out that increases in employment (a robust finding for the EITC for single mothers) could independently lead to a decrease in smoking (HMS p.178). In the end, changes in income alone are not the only channel for the effect of the EITC on smoking. Thus, as we discuss in HMS, our approach does not presume or rely on smoking being an inferior good.

There are three papers that focus on the effects of the EITC on smoking, and find results consistent with ours. Averett and Wang (2013) use several difference-in-differences designs to show that smoking declines among low education mothers after the OBRA93 expansion. Cowan and Tefft (2012), similarly use difference-in-differences around the OBRA93 expansion to show that smoking declined among unmarried women with less than a college degree. In contrast Kenkel, Schmeiser, and Urban (2014) leverage changes to the maximum state and federal credit over time to show that smoking increases among low income adults. While no paper we know of has resolved the differences between these papers, the pattern of findings fits our hypothesis: work focusing on mothers or unmarried women find declines in smoking consistent with increases in employment among these groups. Kenkel, Schmeiser, and Urban find increases in smoking when they focus on the population of adults, potentially driven by men who do not change their labor supply in response to the EITC (Kenkel, Schmeiser, and Urban do not present separate results for men and women smokers).

It is also important to recognize that pregnant women have a unique relationship with smoking due to the role that information and prenatal care play. Elevating women out of poverty could change their information set, the quality of prenatal care they receive, or their decision to act on medical advice related to smoking. These changes could occur in a variety of ways that are not

yet understood by the literature and that we are unable to measure, but suggest a more nuanced story than one solely reliant on the income elasticity of smoking. Ultimately we report effects on smoking because that is the evidence provided by our research design. We admit that it is not yet well understood how income, employment, and prenatal smoking interact.

Further DJ in footnote 7 go on to say "*One of the authors (Hoynes) has noted in an analysis of the food stamp program that smoking among low income women could be a normal good 'The increased transfer income could also encourage behaviors that could harm infant health such as smoking and drinking.' (Almond Hoynes and Schanzenbach 2011 p 391)*."

The food stamp program provides an income transfer that discourages work (Hoynes and Schanzenbach 2012). This is the opposite labor supply response to the EITC. If the effects on smoking operate through both income effects and employment effects then we would not expect that the effects of the EITC and food stamps on smoking to be the same. Further, the time period looked at in Almond Hoynes and Schanzenbach 2011 reflected a period where more pregnant women smoked and less was understood about the consequences of smoking during pregnancy.

6) [DJ Section V, p. 11] DJ argue that we wouldn't expect to see an effect anyways, given the medical literature.

DJ argue that based on the medical literature we would not expect to see effects of the EITC on low birthweight. They cite a 2007 review by the Institute of Medicine (IOM) which states in the abstract: "The current methods for the diagnosis and treatment of preterm labor are currently based on an inadequate literature, and little is known about how preterm birth can be prevented."

We agree that there is a lack of understanding of biological mechanisms for individual cases of preterm birth, this is reflected in a low R^2 in many models of preterm birth. However, that does not mean that there are not plausible interventions that can change individual behavioral and environmental factors that influence the occurrence of pre-term birth. This is also recognized by the medical community. The same abstract cited by DJ from the IOM states directly before their quotation: "Preterm birth is a complex cluster of problems with a set of overlapping factors of influence. Its causes may include individual-level behavioral and psychosocial factors, neighborhood characteristics, environmental exposures…" (Institute of Medicine, 2007).

Relatedly stress is a documented predictor of pre-term birth (Berkowitz and Papiernik, 1993). There is a small but growing literature documenting that increases in income can lead to reductions in stress (Aizer et al 2016, Evans and Garthwaite 2014, Fernald and Gunnar 2009 and Haushofer and Shapiro 2016). Of particular note in this context is Evans and Garthwaite (2014) who find that increases in the EITC lead to reductions in maternal stress.

Additionally, we note the many misstatements in DJ:

- DJ state on p. 2 that "*Our re-analysis underscores the difficulty of estimating the causal effect of a national policy that lacks variation across time and space.*" This statement misses the fact that there are tax law changes in 1986, 1990, 1993, 2001 and 2009) that create variations over time.

- DJ model (2) p. 3 is not correct. In HMS p. 184 we specify our base model for the analysis of multiple expansions. DJ model (2) includes ρ_i – this is parity? If so, it should have a p subscript consistent with the dependent variable. It also does not include the fixed effects for state. Additionally, we include parity by linear time in robustness but it is not in the main model (so we don't include it in HMS equation 2).

- On page 5 DJ state "*five years after the 1993 expansion for women with no previous live births (parity) relative to women having a second or higher order birth (parity 2+)*" And again "*HMS report no association between the 1993 EITC and lbw among white or Hispanic women of parity 2 relative to parity 3+.*" It is the reverse: HMS examine births for women with 2+ (the treatment) relative to first births (control) or parity 3+ (the treatment) relative to parity 2 (the control).

- On the second paragraph of page 8 in DJ "*declines in low birth*" should be "declines in low birth weight"

- On page 10, in describing their standard error correction, DJ describe the procedure results such that the "*point estimates are almost identical*." This procedure should result in point estimates that are identical.

- On page 11 DJ state that "*HMS report that the EITC is protective against very small changes in preterm birth but warn that gestational age is not well-measured and not reported for some state/year cells (HMS's table 6). They also report that EITC lowers the percent of births less than 2000 grams by approximately 1 percent. Birth Weight is well measured. In 2015, 94 percent of all births less than 2000 grams were preterm.*" The issue that we refer to here is that gestational age in the birth certificate data is based on self-reported date since last menstrual period and therefore not well measured.

- On page 11 DJ state that the EITC is "*delivered primarily in February and spent soon thereafter on mostly durables and transportation.*" Patel (2012) finds that the EITC leads to increases in primarily non-durables, work related and housing expenditures. Increases in durable and investment expenditures are concentrated in the first quarter, around the time of refund receipt. But taking account of the changes throughout the rest of the year, the effects are more concentrated in non-durables.

- On page 13 DJ state "*The increases in standard errors as we aggregated the data to the level on the intervention most likely reflects the inherent uncertainty in the exercise*." However, as we discuss above, the standard errors they present are not consistently estimated.

- In the abstract DJ state "*We eliminate any associations between multiple expansions of the EITC and low birth weight across race and ethnicity with simple controls for trends in parity.*" And similarly, on p. 5, "*The EITC may have improved infant health among subgroups of participants,*

but the absence of an association for any group with relatively simple controls for trend should weaken confidence in interpreting HMS's results as causal." This is only true of quadratic trends, not linear trends. Similarly, in the text DJ should clarify that when they refer to "controlling for trends" they mean quadratic trends such is in line 10 on page 19. Finally, as discussed above, whether these controls are "simple" is not a good criteria for evaluating their impact on identification and model estimates.

- DJ section 5 focuses on "pre-term births". However, our paper is about Low birthweight (LBW) status and not preterm birth. While the two are correlated, they are not the same. By moving between discussing our LBW results and citing medical literature that references preterm births, the distinction between the two is conflated and it leads to the impression that clinical research that applies to preterm births applies equally to low birth weight. This distinction warrants greater clarity in this section.

References

Aizer, Anna, Stroud, Laura and Stephen Buka (2016) "Maternal Stress and Child Outcomes: Evidence from Siblings" Journal of Human Resources, 51(3): 523-555.

Averett and Wang (2013). "The Effects of Earned Income Tax Credit Payment Expansion on Maternal Smoking," Health Economics, 22: 1344–1359.

Bitler, Marianne, Hilary Hoynes and Elira Kuka (2017). "Do In-Work Tax Credits Serve as a Safety Net?" *Journal of Human Resources* Vol 36, Issue 2, pp. 358-389.

Cowan Ben, and N. Tefft (2012). "The Effect of Earned Income Tax Credit Expansions on the Smoking Behavior of Women". The B.E. Journal of Economic Analysis & Policy.

Evans, William N., and Craig L. Garthwaite. 2010. "Giving Mom a Break: The Impact of Higher EITC Payments on Maternal Health." American Economic Journal: Economic Policy, Vol 6, no. 2: 258–90.

Fernald, Lia and Megan Gunnar 2009. "Effects of a poverty-alleviation intervention on salivary cortisol in very low-income children," Soc Sci Med. 2009 Jun; 68(12): 2180–2189.

Haushofer, Johannes and Jeremy Shapiro (2016). The Short-Term Impact Of Unconditional Cash Transfers To The Poor: Experimental Evidence From Kenya.

Hoynes, Hilary and Ankur Patel (2018). "Effective Policy for Reducing Inequality? The Earned Income Tax Credit and the Distribution of Income," Forthcoming, *Journal of Human Resources*.

Hoynes, Hilary and Diane Whitmore Schanzenbach, 2012. "Work Incentives and the Food Stamp Program," *Journal of Public Economics* 96(1-2): 151-162.

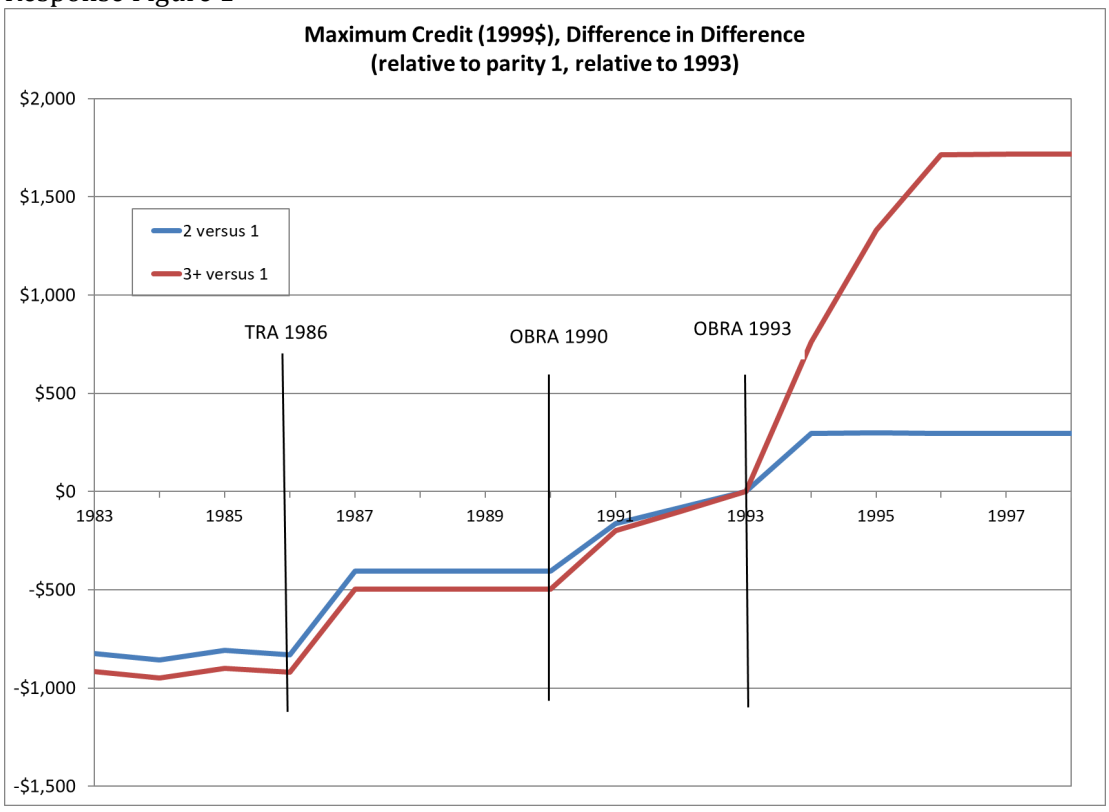
Kearney, Melissa and Phillip Levine (2015). "Investigating recent trends in the U.S. teen birth rate," Journal of Health Economics, 41(2015):15-29.

Kenkel, Schmeiser, and Urban (2014). "Is Smoking Inferior? Evidence from Variation in the Earned Income Tax Credit," Journal of Human Resources, Volume 49, Number 4, Fall 2014, pp. 1094-1120.

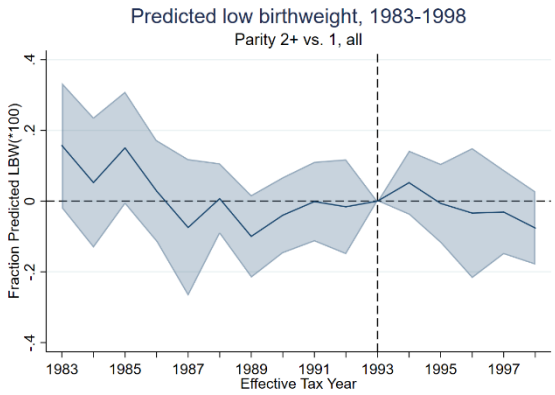
Meyer, Bruce D., and Dan T. Rosenbaum. 2001. "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers." Quarterly Journal of Economics 116, no. 3: 1063–1114.

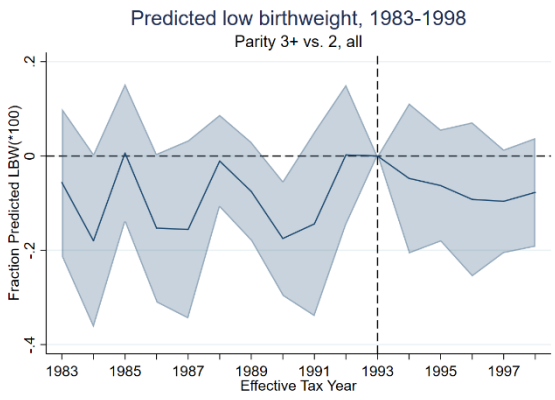
Patel, Ankur (2012). "The Earned Income Tax Credit and Expenditures", U.S. Department of Treasury, November 2012.

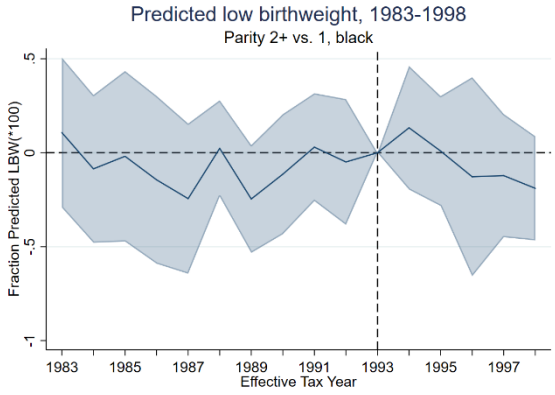
Response Figure 1

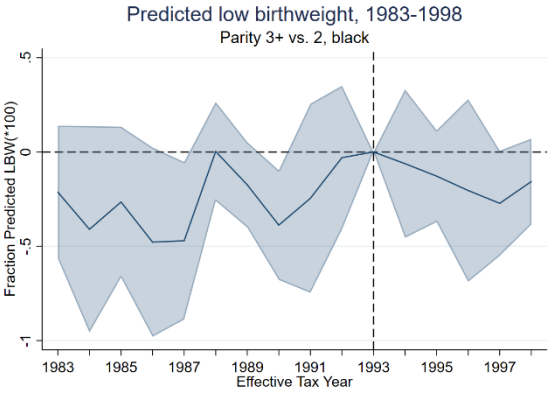


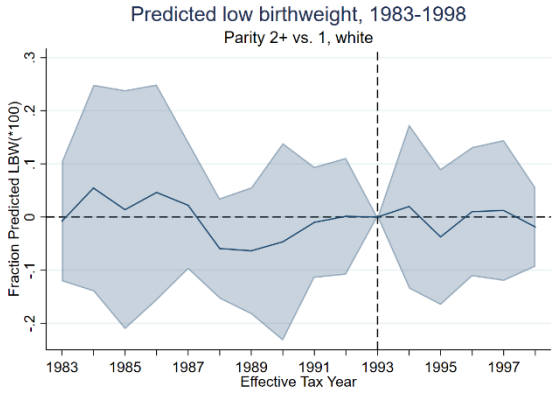
Response Figure 2

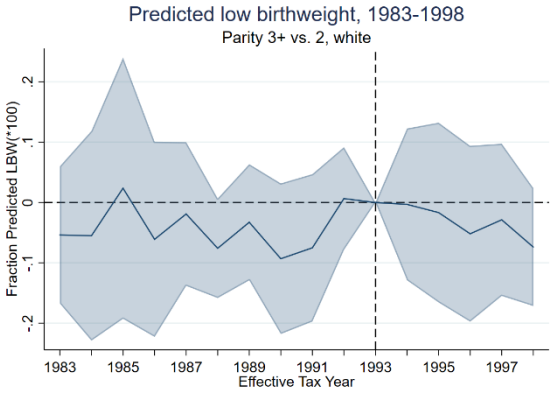




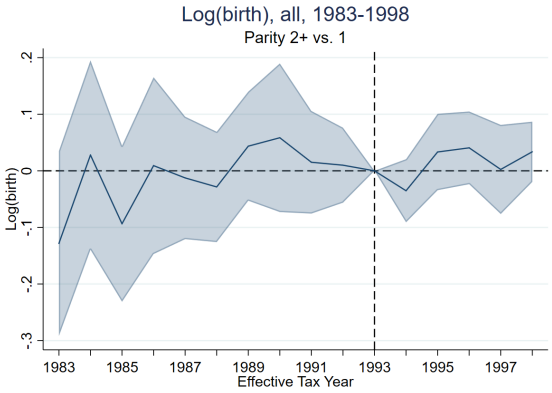


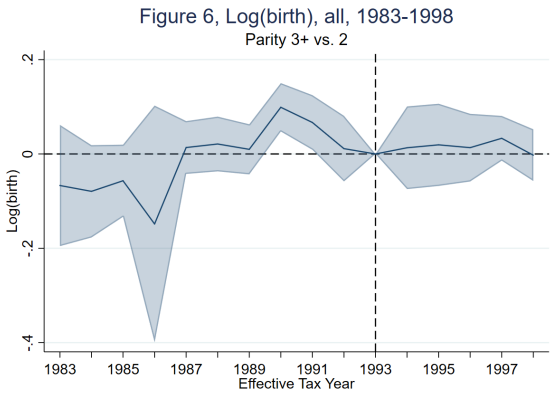


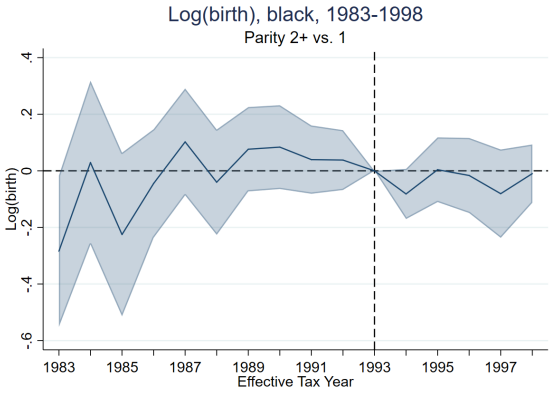


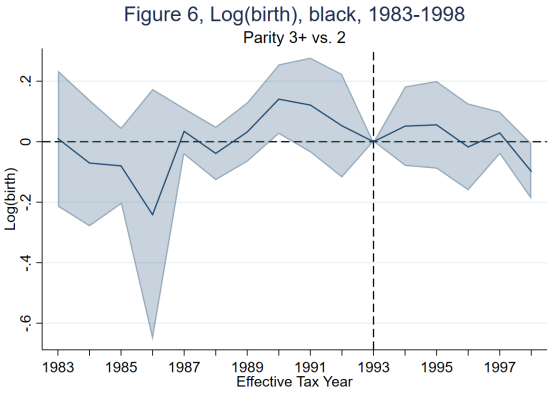


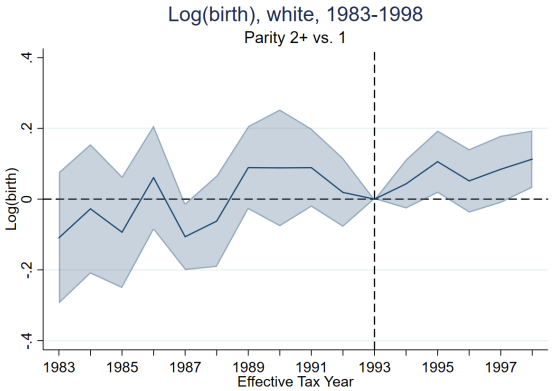
Response Figure 3

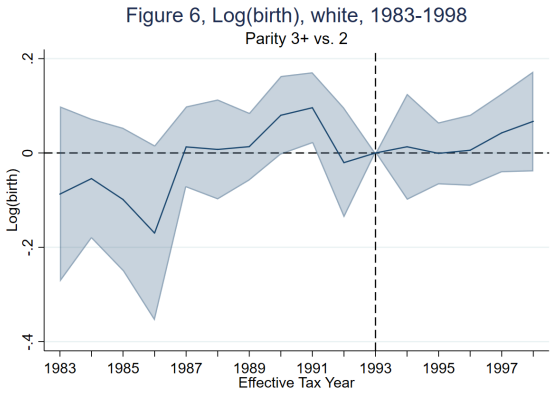




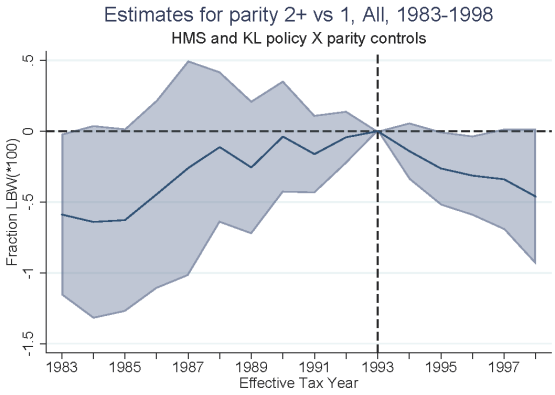


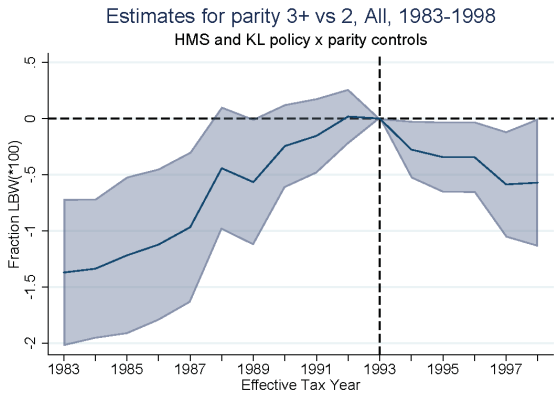


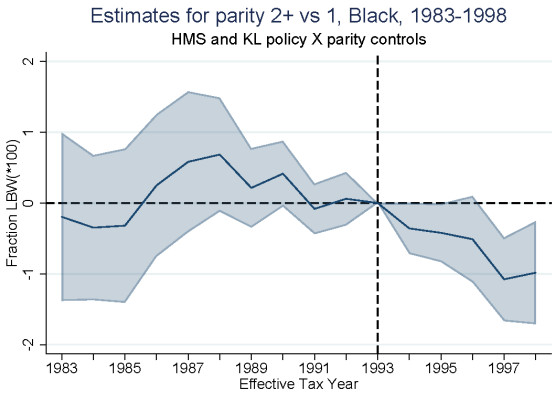


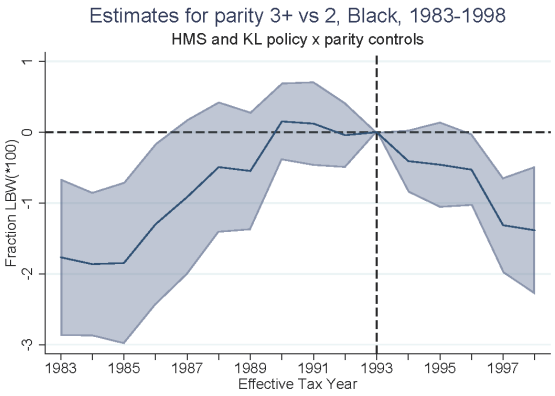


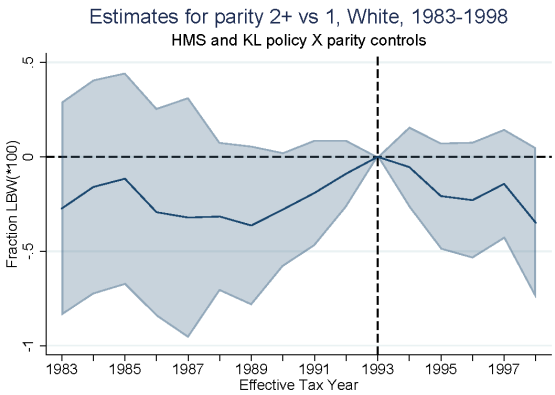
Response Figure 4

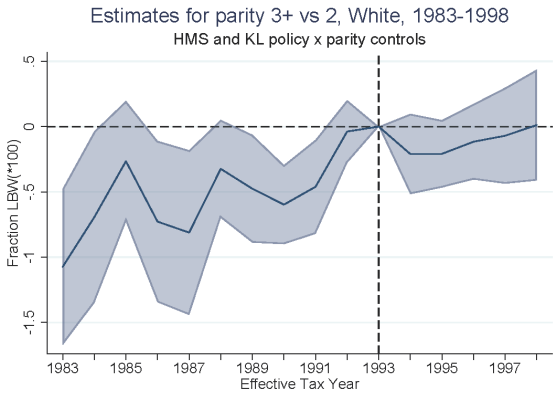




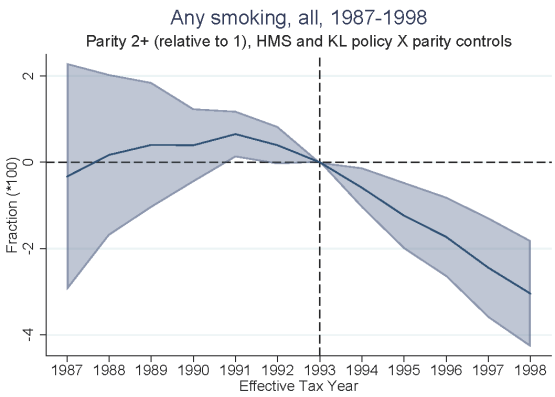


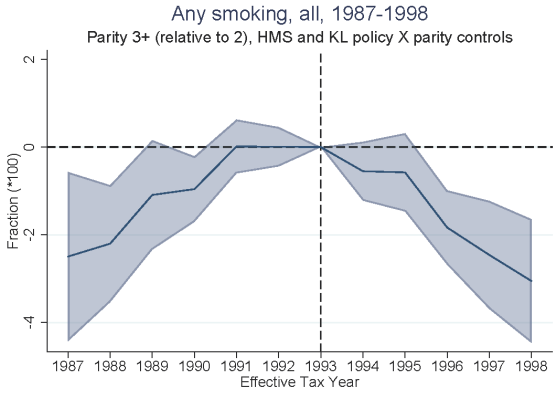


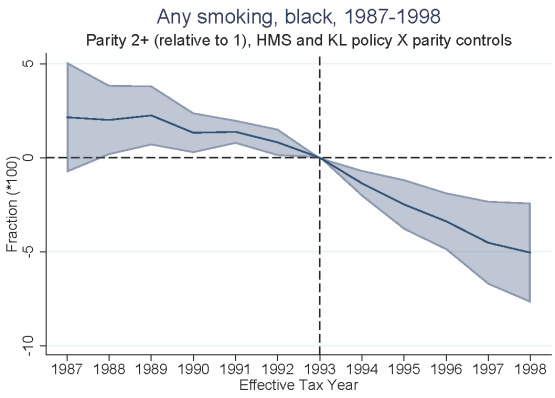


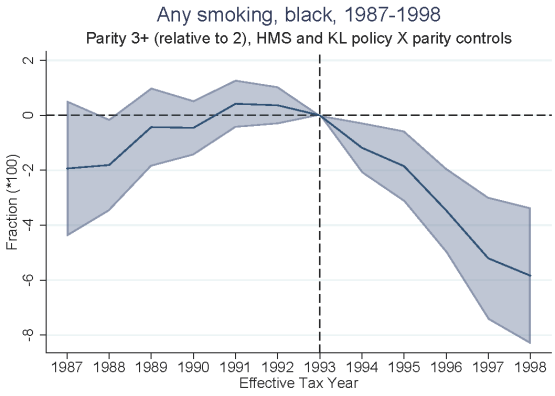


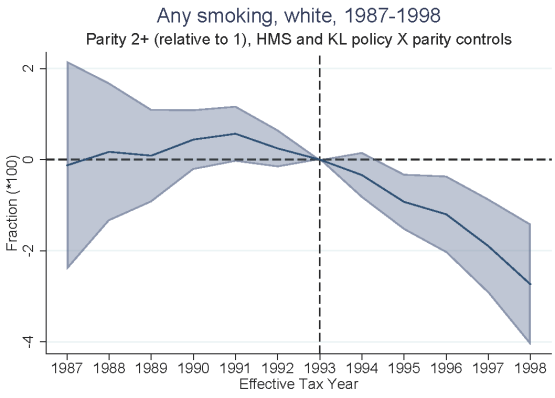
Response Figure 5

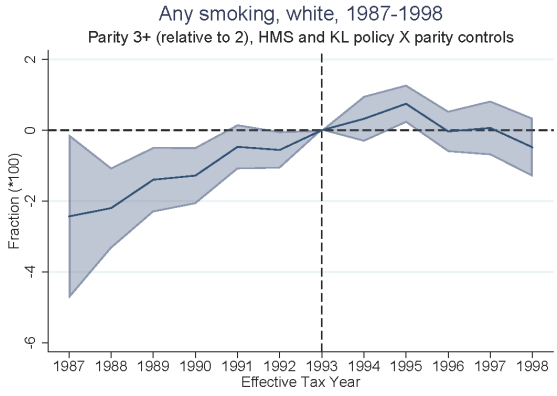




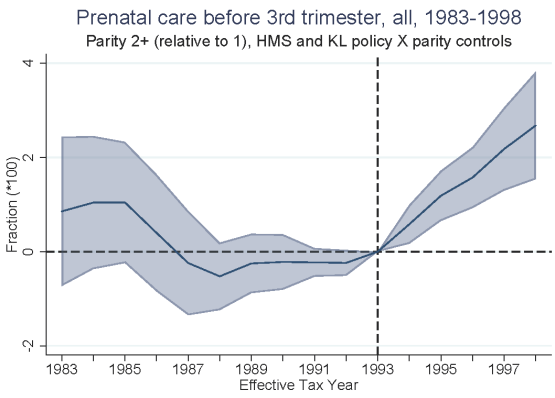


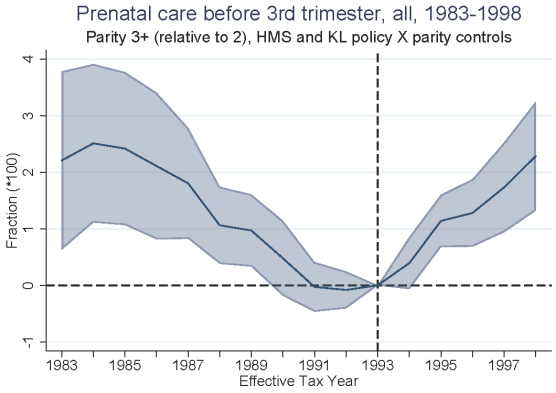


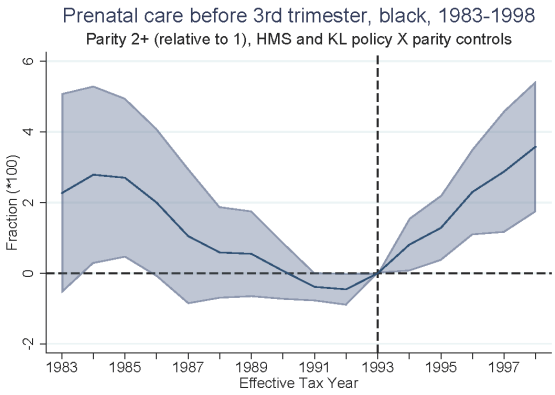


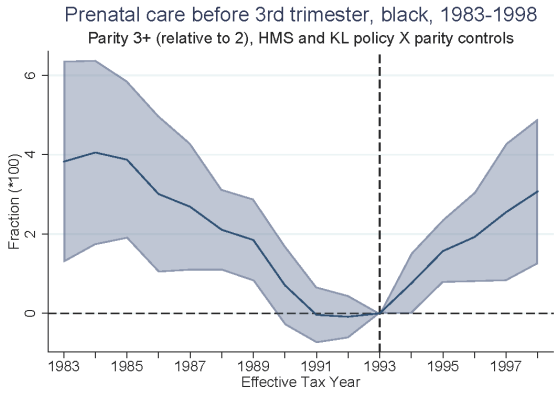


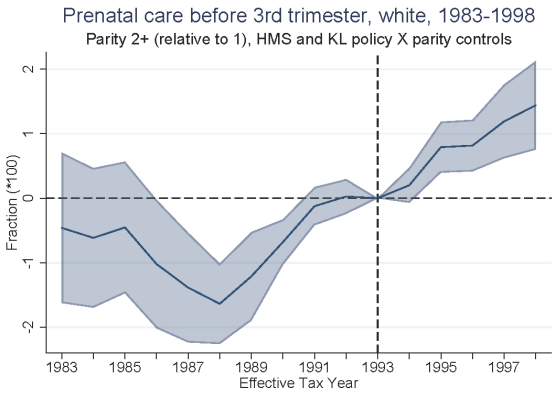
Response Figure 6

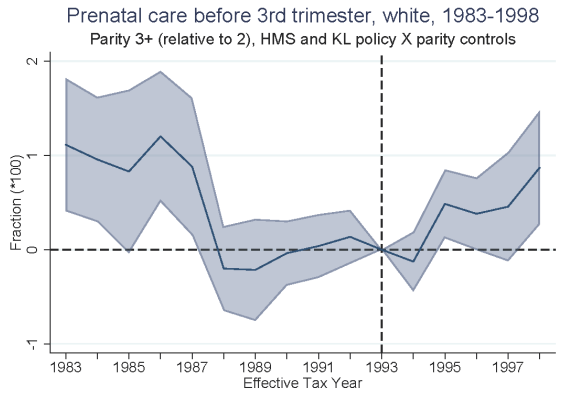




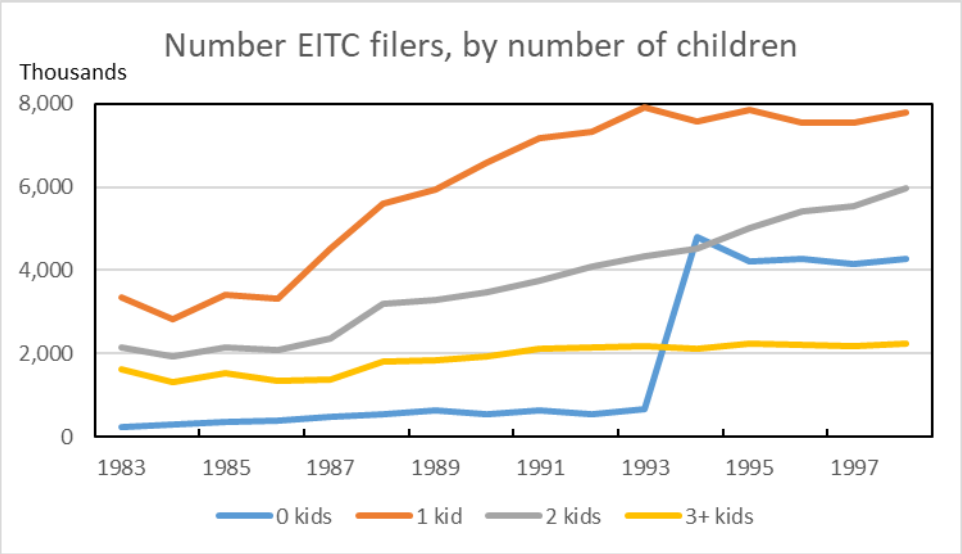


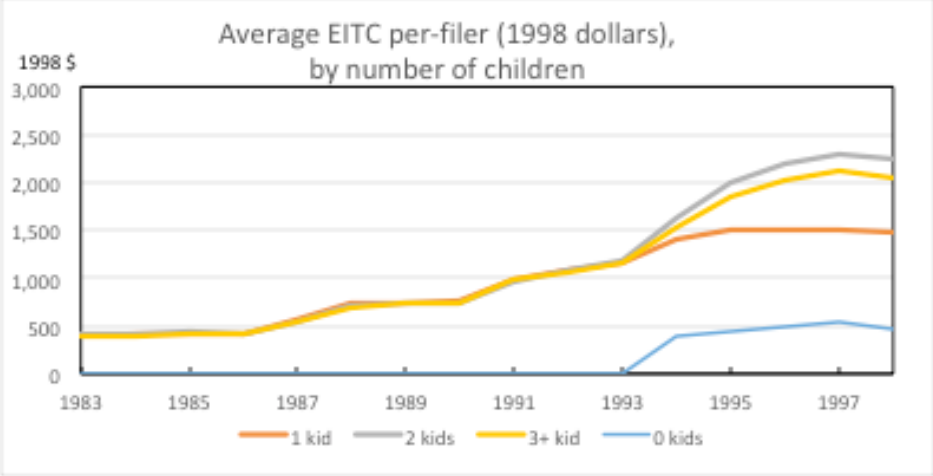






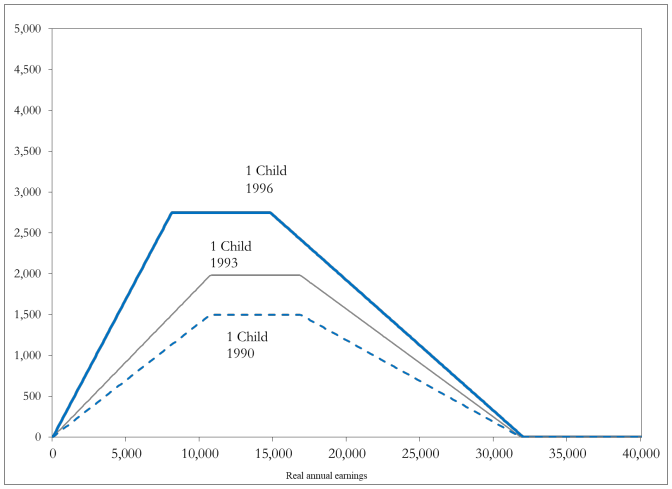
Response Figure 7: Number of EITC filers and average EITC, by number of EITC children



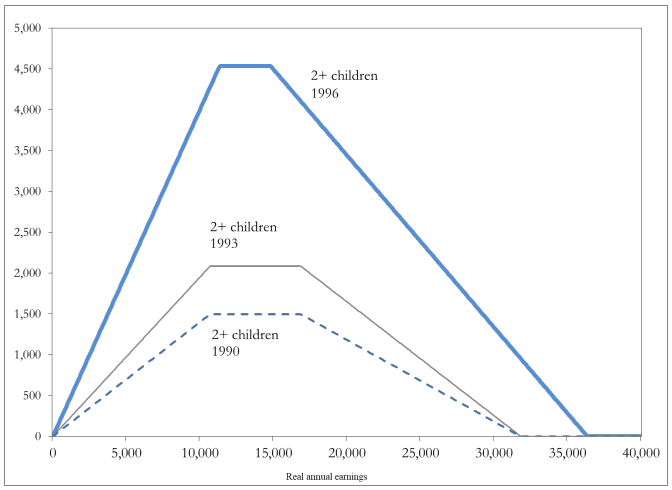


Response Figure 8: EITC Schedule, by family size (real 2006 dollars)

A: 1 child filers



B: 2+ child filers



Response Table 1: Maximum Credit (1999$) by family size (parity of birth)

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Difference (relative to 1993)												
childless (parity 1)	$0	$0	$0	$0	$0	$0	$0	$347	$346	$346	$347	$347
1 child (parity 2)	-$407	-$407	-$406	-$406	-$163	-$81	$0	$642	$646	$644	$643	$644
2+ children (parity 3+)	-$497	-$497	-$496	-$496	-$199	-$99	$0	$1,106	$1,677	$2,061	$2,063	$2,063
Difference in difference (across parity, relative to 1993)												
2 vs 1	-$407	-$407	-$406	-$406	-$163	-$81	$0	$296	$300	$297	$296	$297
3+ vs 1	-$497	-$497	-$496	-$496	-$199	-$99	$0	$760	$1,331	$1,714	$1,717	$1,716
3+ vs 2	-$90	-$90	-$90	-$90	-$36	-$18	$0	$464	$1,031	$1,417	$1,421	$1,420

Response Table 2 (reproduced from HMS response document)
Main estimates using alternative clustering assumptions

Cell level:	Baseline		Micro data		Baseline		Baseline		Baseline		Parity x Year x Demog Group		Parity x Year x Demog Group	
Cluster level:	Baseline (state)		Baseline (state)		Demographic group		Parity x Year		Two-way: Demog group and State		Parity x Year x demog group		Parity x Year	
Parity2+ * After	-0.354***		-0.354***		-0.354**		-0.354***		-0.354***		-0.351**		-0.351***	
	(0.074)		(0.074)		(0.128)		(0.068)		(0.132)		(0.132)		(0.068)	
Parity=2 * After		-0.164**		-0.164**		-0.164**		-0.164***		-0.164**		-0.155**		-0.155***
		(0.072)		(0.072)		(0.058)		(0.042)		(0.072)		(0.062)		(0.040)
Parity 3+ * After		-0.528***		-0.528***		-0.528**		-0.528***		-0.528***		-0.530**		-0.530***
		(0.090)		(0.090)		(0.193)		(0.067)		(0.191)		(0.197)		(0.063)
N	47,687	47,687	6,672,030	6,672,030	47,687	47,687	47,687	47,687	47,687	47,687	1,715	1,715	1,715	1,715
R-sq	0.62	0.62	0.02	0.02	0.62	0.62	0.62	0.62	0.62	0.62	0.93	0.93	0.93	0.93

Response Table 3: Clustering on year

Cell level:	Baseline			Microdata			Group X Parity X Year			Parity X Year		
Cluster level:		Year			Year			Year			Year	
Parity2+ * After	-0.354**			-0.354**			-0.351**			-0.357**		
	(0.0722)			(0.0721)			(0.0677)			(0.0890)		
Parity=2 * After			-0.164**			-0.164**			-0.155**			-0.162**
			(0.0498)			(0.0498)			(0.0466)			(0.0628)
Parity 3+ * After		-0.340***	-0.528***		-0.340***	-0.528***		-0.352***	-0.530***		-0.360**	-0.535**
		(0.0532)	(0.0957)		(0.0531)	(0.0956)		(0.0520)	(0.0903)		(0.0710)	(0.121)
N	47687	35467	47687	6672030	3953134	6672030	1715	1288	1715	32	24	32
R-sq	0.6174	0.6232	0.6176	0.0157	0.0187	0.0157	0.9333	0.9494	0.9336	0.9784	0.9921	0.9863