Does Reality TV Induce Real Effects?
A Response to Jaeger, Joyce, and Kaestner (2016)

Melissa S. Kearney (University of Maryland) and Phillip B. Levine (Wellesley College)[1]

October 31, 2016

I. Introduction

We are grateful to David Jaeger, Ted Joyce, and Robert Kaestner for taking the time and care to undertake a replication and reassessment of our 2015 paper "Media Influences on Social Outcomes: The Impact of MTV's *16 and Pregnant* on Teen Childbearing." We agree with the authors that replication is an extremely important and valuable exercise.[2] We further agree that because there are potential policy implications of our findings, a reexamination of the data and analysis is especially worthwhile. In this paper we respond to their critique paper distributed as IZA discussion paper 10317 in October 2016.[3]

To reiterate what we did and found in our 2015 paper, our main analysis relates geographic variation in changes in teen childbearing rates to viewership of the show. We implement an instrumental variables (IV) strategy using local area MTV ratings data from a pre-period to predict local area *16 and Pregnant* ratings. We find that in places where more teens are predicted to be watching the show, teen childbearing rates fell by more than areas with lower

[1] *Acknowledgements:* We are grateful to Kristin Butcher, Lisa Dettling, Dan Fetter, Judy Hellerstein, Robin McKnight, and Fernando Saltiel for their thoughtful comments on an earlier draft. Of course, all responsibility for the content of this paper remains our own.
[2] The issue of an inability to replicate results in social science research has received considerable attention in recent years. See, for instance, Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." *Science*, 349(6251). Available at: http://science.sciencemag.org/content/349/6251/aac4716 (accessed 10/28/2016).
[3] Available at http://ftp.iza.org/dp10317.pdf.

rates of predicted viewership after the show was introduced. The results imply that the MTV show led to a 4.3 percent reduction in teen births.

The primary objective of a replication exercise is to conduct exactly the same exercise and verify that the results have been produced properly. We are pleased that Jaeger, Joyce, and Kaestner (henceforth JJK) were able to reproduce every single finding in our paper.[4] That is to say, after careful examination of our data and programs by JJK, the results reported in Kearney and Levine 2015 (henceforth KL) are maintained. We view this as the crucial outcome of their detailed replication efforts.

The basis of their critique of our work is thus not about failure to replicate, but rather about the econometric issue of parallel trends.[5] Discussions of parallel trends usually relate to a simple difference-in-differences approach, but the concept is relevant in the context of our IV strategy as well, which is easily seen when considering the reduced form specification. In that framework, an important assumption underlying our identification strategy is that teen birth rates

[4] JJK did identify two typos. In JJK footnote 10, they report that the notes to Figure 5 of KL failed to indicate that the specification included the DMA (designated market area – i.e. TV market) unemployment rate, though the description in the text explains precisely what is in the model. Once they included that variable in the model, their results were identical to those we reported. They also note in JJK footnote 13 that the p-value for the joint test of significance of all the pre-period interactions with MTV ratings is 0.13, not 0.21 as the paper states.

[5] One other criticism that JJK make (on page 2) is that "even KL concede that their instrument does not provide a truly exogenous source of identifying variation of the viewing of *16 and Pregnant*." We do not concede that point. Our identifying variation comes from different rates of MTV viewership *before* the show began. Since those ratings are determined prior to June 2009, our instrumental variable strips out the variation in *16 and Pregnant* ratings that are specifically attributable to that particular show, and potentially reflective of a time-varying latent preference for a show about teen mothers. We do acknowledge in KL that using this source of variation means that we are estimating an effect relevant to the MTV viewing population of teens, not a "random" teen. This allows us to ask the question of what this show achieved, but it does not answer the different question about how the show would affect teen birth rates if a "random" teen were assigned to viewership. This is not a threat to identification of the causal relationship we are interested in.

would not have experienced a larger relative drop in television markets with relatively higher pre-period MTV ratings immediately following June 2009, were it not for the introduction of *16 and Pregnant*. This can be thought of as an assumption that the birth rates of places with different rates of pre-period MTV viewership were essentially on parallel trends. The thrust of the JJK critique of our paper is an alleged violation of this assumption.[6]

In this response, we provide more detail regarding the parallel trends assumption and how it could affect empirical results. After that, we review the analysis in our original paper, highlighting the fact that we explicitly considered the issue of parallel trends. During the period we study, no violation of this assumption is apparent. We then discuss in detail the major findings from the JJK reassessment. We demonstrate in this paper that JJK's critique does not pose a serious threat to the interpretation of our findings. Our reading of the totality of evidence is that it supports the findings of our 2015 paper.

II. The Parallel Trends Assumption

The violation of the parallel trends assumption would be a threat to identification in our quasi-experimental analysis of the impact of *16 and Pregnant*. It would be a problem for our analysis if teen birth rates in local television markets (called "Designated Market Areas" or DMAs) were trending in ways correlated with MTV viewership rates before the introduction of

[6] The JJK paper also works to demonstrate the fragility of the social media regressions in the KL paper that illustrate spikes in Google Searches and Tweets containing the terms "birth control" after the show was introduced. That analysis was secondary and suggestive, the data was necessarily limited, and we would not be surprised if the results are sensitive to various weighting schemes. As the robustness of those estimates is crucial neither to the finding of a reduction in births nor to the JJK criticism of our main findings, we do not address it in this response. That said, we note that even a non-regression adjusted look through the Twitter data reveals tweet after tweet explicitly saying some version of "watching *16 and Pregnant* makes me want to take birth control." Those data give the clear impression that this show has affected viewers' attitudes.

the show. Again, for simplicity, we focus this discussion on the reduced form version of our analysis. If DMAs with higher MTV ratings were already experiencing a more rapid decline in teen birth rates relative to those DMAs with lower MTV ratings before the show was introduced, those differential trends could have continued beyond the show's introduction. A comparison of teen birth rates before and after the introduction of the show would then indicate that teen births were lower in the locations with higher MTV ratings, but it would not necessarily reflect a causal effect of the show.

Empirical researchers utilizing quasi-experimental variation routinely use multiple approaches to address the question of whether the parallel trends assumption is appropriate in a given context. A preliminary approach typically involves visual examination of the data. In our context, we could compare the locations where ratings were higher to locations where ratings were lower, plot trends in teen birth rates before and after the intervention, and visually inspect whether differential trends were occurring beforehand. We could take this approach one step further and estimate econometric models testing whether any observed differences are statistically meaningful or whether they could have plausibly occurred by chance. An alternative approach would be to introduce a "placebo test" where we suppose that the intervention occurred at a different time than it actually did. If parallel trends were violated, the data would indicate a decline in teen birth rates after the simulated intervention date even though no such causal response could have been possible. A potential solution to this problem, should it exist, is to econometrically control for the differences in pre-existing trends and then determine whether a discrete decline in teen births occurs after controlling for them.

III. KL Discussion of Parallel Trends

In our 2015 paper (KL), we devote considerable attention to the issue of parallel trends. Our discussion of the issue starts on page 3613. We begin with visual inspection of the data and augment that with statistical tests. The relevant text and figure is included here for convenience.[7]

> *In Figure 5, we present a simplified, reduced form version of Equation [5], where quarterly indicator variables are interacted with 2008-09 MTV ratings (the instrument in our IV model) in a model that only includes quarter and DMA*season fixed effects to absorb much of the residual variation. Figure 5 simply plots these quarterly interaction coefficient estimates along with their confidence intervals.[8] For the period before the show began, coefficients are not statistically significant (with one exception among 14 coefficients). Jointly, the group is not significantly different from zero (p-value = 0.21). After the show began, all the coefficients turn negative and two of the six are statistically significant. A test of joint significance of those six coefficients rejects that they are all equal to zero (p-value = 0.002).*
>
> *Still, visual examination of the pre-period suggests a downward slope that goes against a causal interpretation, despite the fact that the F-tests indicate that the higher values during that period are insignificant. To take this visual analysis one step further, we simply introduce best fitting lines through the pre- and post-period coefficients, which are also shown in the figure. These lines make it clear that there is, indeed, a downward slope before 16 and Pregnant began, but there is still a sizeable downward shift in point estimates precisely at the point when the show was introduced. We believe this provides strong visual support for the notion of a causal effect.*

[7] As we indicated in footnote 4, the p-value statistic reported in the first paragraph below should be 0.13, not 0.21, as our original paper stated.

[8] In a model with DMA*season fixed effects, the base period of omitted interactions totals four, one for a particular DMA in each of the four seasons. This means that the base period is one year long, not one quarter long. We have chosen to use July 2008 through June 2009 as the base period to make it easier to see the "treatment effect," which occurs once the show starts in June 2009.

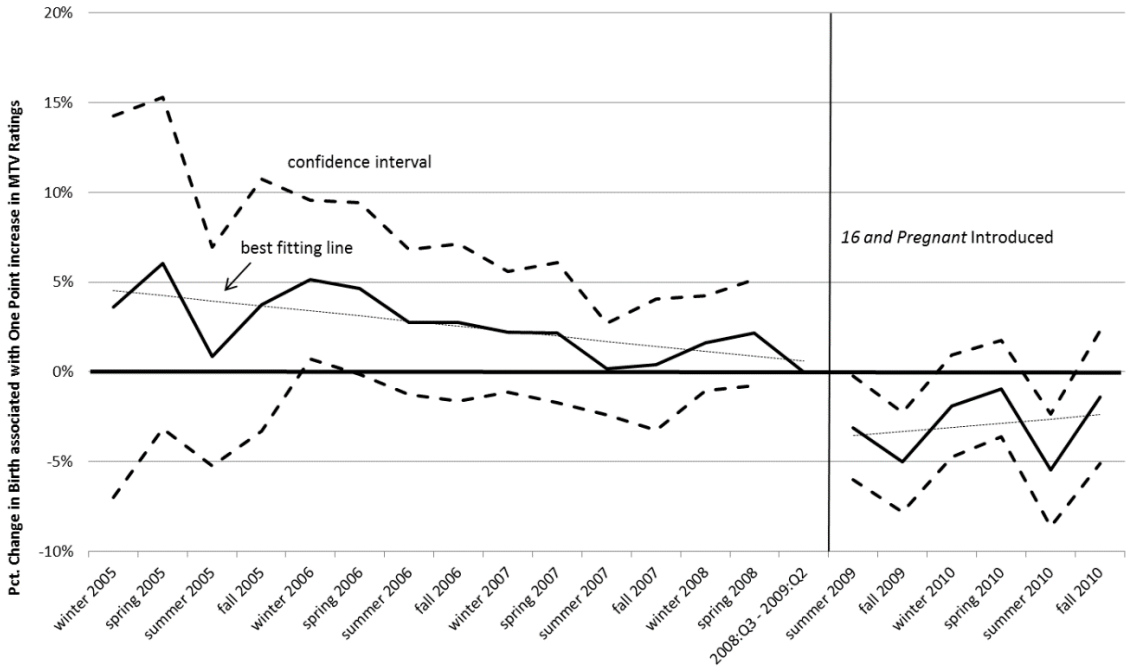


Figure 5: Reduced Form Event Study Estimates of the Impact of *16 and Pregnant*

notes: Estimates reflect coefficients on 2008-09 MTV Ratings interacted with Quarter from a regression model controlling for DMA*season fixed effects. Dashed lines reflect 95% confidence intervals. Dotted lines reflect the best fitting lines through the pre- and post-period coefficients.

In KL Appendix Table B1, we reported the results of a formal event study, plotting out the coefficients of interactions between MTV ratings and quarterly dummy variables in a specification that is otherwise the same as our main findings in the paper, reported in Table 1. We have included that table as an appendix to this response. The econometric evidence shows a sharp and discrete decline in teen birth rates that occur exactly at the point that *16 and Pregnant* began.

Our original paper did not include a formal placebo test, but we report one here. The placebo test we estimate and report uses data from our analysis sample window. This is arguably the most appropriate placebo test to run. The data from our actual analysis represents the 24 quarters that begin in 2005:Q1 and the intervention occurs in 2009:Q3. For our placebo test we use the same data eliminating those quarters after the intervention and simulate the impact of an intervention occurring in 2008:Q1 with 6 quarters after the intervention, as in our actual analysis.

The results of this analysis indicate a statistically insignificant impact of the simulated intervention. The main reduced form estimate from Table 1 in KL is a coefficient (standard error) on 2008-09 MTV ratings of -3.581 (1.512). In our placebo test specification, we get -1.314 (1.696). This placebo analysis provides no indication of non-parallel trends.

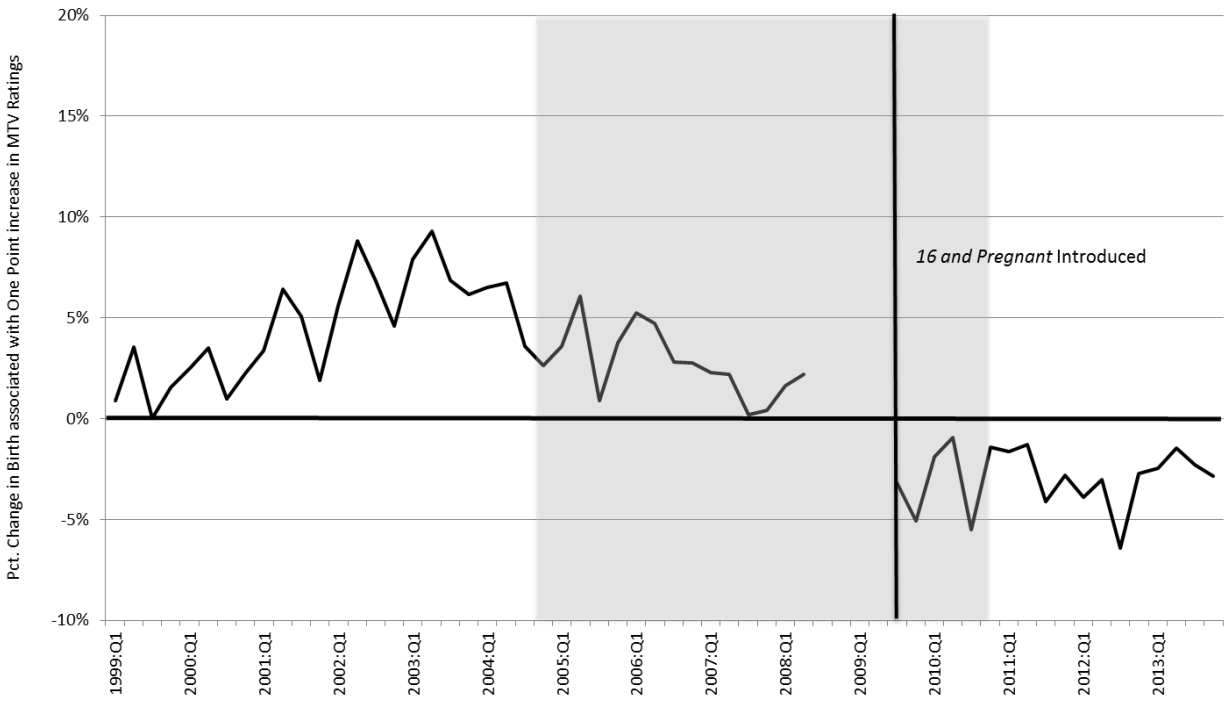
What we conclude from all of this analysis is that during our sample window, there is no indication of potential differential trends across television markets with different levels of MTV ratings before the show was introduced. This supports our interpretation that the estimated effect reflects a causal effect of *16 and Pregnant*.

IV. JJK: Parallel Trends over Longer Sample Windows

What JJK contribute to this line of investigation is to extend the data further backwards, including years prior to 2005.[9] Figure 5 in their paper (hereafter referenced as JJK Figure 5) is the easiest way to understand the impact of extending the sample window. We report JJK Figure 5 here, modified slightly to incorporate minor aesthetic changes.[10] As presented, this figure is directly comparable to KL Figure 5 except that it uses data from 1999 through 2013. The shaded region represents the years used in KL Figure 5.

[9] JJK also extend the sample window forward, but the issue of parallel trends depends on patterns in the data in the period before the intervention occurs. Additional years after treatment is helpful for observing whether the initial reduction in birth rates is maintained. According to JJK Figure 5, that appears to be true.

[10] We are grateful to the authors for providing us with the specific values used to recreate JJK Figure 5. We focus on Panel A of JJK Figure 5 because it treats the period just before the introduction of the show as the base period. We also incorporate JJK's display of the full four quarters of the base period, rather than compressing them to one "tick" on the X axis as KL do. We also omit the confidence intervals to simplify the image. The Y-axis scale of this figure has also been modified to be consistent with KL to facilitate direct comparisons. Aside from these modifications, this figure is substantively identical to JJK Figure 5, Panel A.



JJK Figure 5: Reduced Form Event Study, 1999-2013

note: The shaded region is the same used in Kearney and Levine (2015).

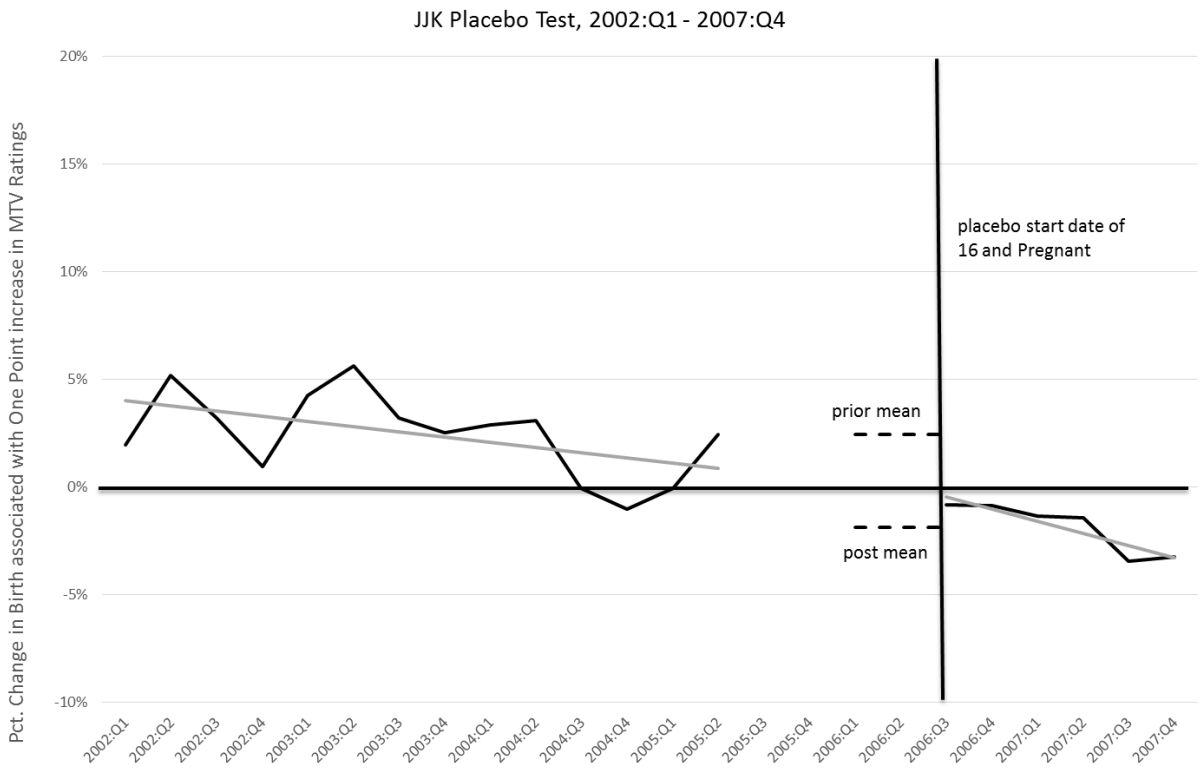
What do we learn from the introduction of these additional years of data? First, we see that extending the sample window all the way back to 1999, all of the estimated differences between places with higher and lower MTV ratings are above the zero-line before 2009 (when the show was introduced), and all of the estimated differences are below the zero-line in the latter period. This is suggestive of a discrete break in trend at the time of the show's introduction. JJK document this point econometrically. In their Table 4, they estimate exactly our model using quarterly data from 1999 through 2013. Their results based on that specification indicate a negative and statistically significant impact of *16 and Pregnant* on teen births that is actually larger than the one we reported.

Second, we see that as JJK point out, if we extend the window back sufficiently far, there appears to be a violation of the parallel trends assumption. That is not necessarily surprising. In

any quasi-experimental analysis, it is quite possible that the parallel trends assumption will not hold if you go back in time far enough. We return to this issue in the conclusion below.

This figure can largely be used to explain other results that JJK present. The first core element of their critique is that placebo tests using data from earlier periods incorrectly identify an impact. Consider JJK Table 1. This table potentially gives a misleading impression of many failed placebo tests. First, all the reported placebo tests in the gray shaded areas of Table 1 should be considered invalid since they include part of the actual treatment. Those are not clean placebo tests. Second, the overlapping nature of the samples in this analysis mean that the rows do not reflect independent tests. JJK introduce rolling 24 quarters windows, imposing a placebo "treatment" 18 quarters into the window. Estimates in each row of that table use almost all the same data as the row preceding it or following it.

A more appropriate way to conduct multiple placebo tests is to focus on exclusively-defined broader periods. Doing that yields a different picture than that presented by JJK Table 1. Placebo tests would pass (that is, there would be no indication of a spurious treatment effect) during the 1999:Q1 through 2004:Q4 period. However, placebo tests conducted during the window between 2002:Q1 and 2007:Q4 fail. This is not surprising given our discussion of the data in JJK Figure 5, which indicated that differential trends exist during this period. A difference-in-differences model using those 24 quarters with a break in quarter 18 (2006:Q3) would definitely show a false treatment effect. We show this in the following figure, which uses data from JJK Figure 5 with this narrower sample window. The mean coefficient in the first 18 quarters have a mean coefficient that is higher than the last 6 quarters in JJK Figure 5, and that would be interpreted within this framework as a treatment effect, albeit a spurious one.



notes: This analysis is based on the coefficients reported in JJK Figure 5 (normalized to average 0 in the placebo base period of 2006:Q3 through 2007:Q2).

Yet, there is a clear contrast between the pattern reported in this figure compared to KL Figure 5 reported above. As in that figure, here we also include best-fitting trend lines through the points in the pre- and post-placebo intervention periods. The trend in the pre-period clearly follows directly into the post-period. In KL Figure 5, a noticeable break from trend is present even after factoring in any pre-existing trend. Just because the estimated effect from the KL sample window in KL Figure 5 and the 2002 through 2007 sample window reported here are both negative and significant does not mean that they both are tainted. The causal effect interpretation is valid in our analysis because of the break in trend and the support for the parallel trends assumption during that time period. No such break is observed here, and the parallel trends assumption does not appear to hold over this time period.

We now turn to a discussion of JJK Table 4, which considers the addition of quadratic trends to the model specification applied to various sample windows. Column 1 presents our original specification and Column 2 augments that specification with DMA-specific quadratic trends. This part of the table, reformatted slightly, is reported below. Column 1 of JJK Table 4 shows that our specification is strongly robust to the sample window used. Across all considered sample windows, the results indicate negative and significant effects of *16 and Pregnant*, which are even larger than the ones we find in our sample window. JJK do not report results with linear DMA-specific trends, but inspection of their Figure 5 suggests that it would generate similar results.[11]

Column 2 of JJK Table 4 introduces DMA-specific quadratic trends. This adjusts for non-parallel trends by allowing birth rates in each television market to be trending differently and non-linearly. In the extended sample window that JJK use, such a specification is likely appropriate given the divergence in trends observed in the mid-2000s. With this modified specification, the point estimates are smaller, but they are still statistically significant in three of the four sample windows. The only sample window where the point estimate approaches zero and is statistically insignificant is the 2001 through 2012 sample window. This is the window that would rely most strongly on those middle 2000s years, which JJK Figure 5 and JJK Table 1 revealed to be problematic.

The remaining columns of JJK Table 4 (not reported in the extract below) consider the sensitivity of the results to models where MTV ratings are converted from a continuous measure

[11] We are unable to estimate these models ourselves because currently we do not have access to the restricted birth data from these earlier years. JJK would not be allowed to share their data with us because they are obtained through a confidentiality agreement with NCHS. Given the typical multi-month lag in gaining approval to obtain these data from NCHS, we would not be able to put out a timely response if we waited to complete that process.

to quartile indicator variables. This converts a continuous measure of MTV ratings into four discrete measures of MTV ratings, thereby reducing the identifying variation. Given that there is no econometric basis for this specification, we do not view this as a particularly informative specification check.

Extract from JJK Table 4: Instrumental Variables Estimates of Impact of *16 and Pregnant* on Birth Rates with DMA-Specific or MTV-Quartile-Specific Trends

	Coefficient (standard error) on *16 and Pregnant* Ratings	
Sample Period	KL Specification: No DMA-Specific trends	JJK Specification DMA-Specific Quadratic Trends
	KL Sample Window	
2005:Q1-2010:QIV	-2.368**	-1.591**
	(0.942)	(0.756)
	Alternative Sample Windows	
2003:Q1-2011:QIV	-3.090***	-1.736**
	(1.064)	(0.714)
2001:Q1-2012:QIV	-3.704***	-0.755
	(1.314)	(0.738)
1999:Q1-2013:QIV	-3.561**	-1.354**
	(1.489)	(0.683)

*** indicates significant at the 1 percent level, ** indicates significant at the 5 percent level.

Yet, JJK go on to argue in JJK Table 5 that even models that include quadratic trends do not generate a causal effect. They repeat the same placebo test exercise that they report in JJK Table 1 except this version includes quadratic trends. We report an extract of that table here, omitting all sample periods that include some treatment quarters, which, as we noted above, violates the premise of a placebo test. We also ignore the problem that we highlighted earlier that

rolling windows should not be thought of as independent tests and are not the proper way to conduct such an exercise. Setting that issue aside, we still interpret the results they present differently than they do. That table reports a negative and statistically significant impact of *16 and Pregnant* in the actual year it was introduced and no evidence of a negative and significant effect in any five-year sample window going back to the year 2000. It is not until the placebo test starts in 1999 that it fails.

Extract from JJK Table 5: Placebo Tests of Estimated Reduced Form and Instrumental Variables Impact on Teen Birth Rates Rolling 24 Quarter Periods with Quadratic DMA-Specific Trends

	Dates			Instrumental Variables	
Row	Begin	"Show" Start	End	Coefficient	Std. Err.
(1)	1999:QI	2003:QIII	2004:QIV	-0.207	1.168
(2)	1999:QII	2003:QIV	2005:QI	-0.788	1.460
(3)	1999:QIII	2004:QI	2005:QII	-1.787*	1.082
(4)	1999:QIV	2004:QII	2005:QIII	-1.906**	0.942
(5)	2000:QI	2004:QIII	2005:QIV	-1.576	1.889
(6)	2000:QII	2004:QIV	2006:QI	-1.036	2.084
(7)	2000:QIII	2005:QI	2006:QII	-0.020	1.383
(8)	2000:QIV	2005:QII	2006:QIII	0.700	0.705
(9)	2001:QI	2005:QIII	2006:QIV	0.404	1.177
(10)	2001:QII	2005:QIV	2007:QI	2.196**	1.006
(11)	2001:QIII	2006:QI	2007:QII	1.386	0.859
(12)	2001:QIV	2006:QII	2007:QIII	0.782	1.150
(13)	2002:QI	2006:QIII	2007:QIV	0.700	1.017
(14)	2002:QII	2006:QIV	2008:QI	-0.241	0.994
(15)	2002:QIII	2007:QI	2008:QII	-1.018	0.876
(16)	2002:QIV	2007:QII	2008:QIII	-0.724	0.574
(17)	2003:QI	2007:QIII	2008:QIV	-0.091	0.664
(18)	2003:QII	2007:QIV	2009:QI	0.375	0.539
(25)	2005:QI	2009:QIII	2010:QIV	-1.591**	0.756

V. Conclusions

Amidst the extensive number of tables and figures in the JKK critique paper and this response to it, a key finding must not get lost: JJK were able to replicate all the results of our paper. The substance of the critique is thus not about failure to replicate, but about the proper interpretation of the evidence.

There are a number of important results to come out of their reassessment and our response here. First, there is no evidence of a parallel trends assumption violation during our sample window of 2005 through 2010. Second, results generated during our sample window are robust to the inclusion of DMA-specific quadratic trends. Third, results generated with our specification are robust to alternative sample windows. Fourth, when the sample window is extended back eight years (but not 10 years) from the intervention and quadratic trends are included in the model, the estimated result is attenuated and becomes statistically insignificantly different from zero. This is really the basis for the JJK conclusion that our finding of an effect of this show is not robust to closer scrutiny. Yet our reading of the totality of evidence still leads us to conclude that the data imply a negative causal effect of the show on teen birth rates.

The disagreement between JJK and us on that interpretation raises a crucial issue for applied researchers about the most appropriate way to investigate a question in the data and interpret the weight of evidence. There are no hard and fast rules regarding the appropriate sample frame or how many specifications should be run. One needs to be sufficiently thorough and confirm that baseline findings stand up to a host of possible alternative specifications. This is important to rule out the possibility that the estimated results are simply a peculiar set of findings, arrived at perhaps through an inadvertent selection of a specification that yielded support for some hypothesis. And yet there is the associated risk that if one estimates too many specifications, the research process moves too far in the other direction, finding a peculiar specification that attenuates the estimated effect. The empirical researcher needs to be as thorough as possible without running so many specifications and considering so many windows as to err on the other side of looking for a particular opposite result. At some point, demanding that the data yield a stable estimate under all years and under all specifications stretches the

demands of robustness too far. This is ultimately a judgment call, about which reasonable researchers can disagree.

In our examination of the impact of *16 and Pregnant*, at the time we conducted our research, birth data was only available for 6 quarters after the introduction of the show. Does it really make sense to use a sample window that includes up to 40 quarters beforehand as the basis of comparison? We think our choice of 18 quarters beforehand is appropriate. It is long enough to provide sufficient statistical power without incorporating the myriad other factors that can occur across places over time. This is why we still conclude that our initial paper identified a causal effect of *16 and Pregnant* on teen birth rates. We stand by the results of our 2015 paper.

Appendix Table B1: Estimates of the Impact of *16 and Pregnant* Ratings on Teen Birth Rates

	OLS (1)	IV (2)	Reduced Form (3)
Rating*(Q1:2005-Q3:2005)	-0.043	1.075	1.782
	(1.148)	(1.897)	(3.077)
Rating*(Q4:2005-Q2:2006)	0.246	1.579	2.464
	(0.874)	(1.272)	(2.069)
Rating*(Q3:2006-Q1:2007)	-0.120	0.993	1.551
	(0.580)	(0.942)	(1.508)
Rating*(Q2:2007-Q4:2007)	-0.234	0.072	0.171
	(0.433)	(0.763)	(1.266)
Rating*(Q1:2008-Q2:2008)	-0.272	0.605	0.933
	(0.494)	(0.685)	(1.121)
Base Period: (Q3:2008-Q2:2009)	---	---	---
Rating*(Q3:2009-Q1:2010)	-1.035	-1.880	-2.816
	(0.334)	(0.562)	(0.957)
Rating*(Q2:2010-Q4:2010)	-1.127	-1.510	-2.328
	(0.400)	(0.677)	(1.178)
Unemployment	-1.435	-1.490	-1.524
Rate	(0.405)	(0.384)	(0.442)

Notes: The data used for this analysis represents quarterly birth rates by DMA for conceptions leading to live births between 2005 and 2010. The sample size in each model is 4919 (205 DMAs, 24 quarters, and one observation was dropped because there were no teen births). The dependent variable, the birth rate, is measured in natural logs. The rating variable represents ratings for *16 and Pregnant* in OLS and IV models and MTV ratings in 2008-09 in the reduced form model. Coefficients and standard errors (reported in parentheses) are multiplied by 100. Each model also includes the percentage of a DMA's female teen that are Hispanic and non-Hispanic black along with quarter and DMA*season fixed effects. Regressions are weighted by the relevant sample sizes for each outcome. Reported standard errors are clustered at the DMA level. Bolded results are statistically significant at the 5 percent level.