**PSCI-19-0739**

**Challenging the Link Between Early Childhood Television Exposure and Later Attention Problems: A Multiverse Approach**

**McBee, Brand, & Dixon**

Dear Dr. Roberts,

Thank you for the opportunity to resubmit our paper, "Challenging the Link Between Early Childhood Television Exposure and Later Attention Problems: A Multiverse Analysis" [PSCI-19-0739] to Psychological Science. We have slightly altered the name of our new submission since our revision now includes two multiverse analyses.

This submission includes many more analytic models than the previous version. The manuscript focuses on a description of our approach and a high-level presentation and discussion of the results. Technical details for each of the 848 models, including coefficient tables, model-specific descriptive statistics, plots, and model diagnostic information, is presented on our project’s new Github page (<https://github.com/mcbeem/TVAttention>), which replaces the former OSF project page. This model-specific supporting information can be found under the “Results” component.

Further, we have increased the computational reproducibility of this project by providing a Docker image containing the entire software toolchain. The justification for doing this was made clear when an update to the *mice* package required a re-write of relevant parts of the analysis script during the revision process. An Appendix to this article gives a step-by-step description of how to reproduce our work with Docker.

We believe that this revision is substantially improved in content, scope, clarity, and rigor. We are grateful to the editor and reviewers for their thoughtful comments and suggestions, which we specifically address below.

**As part of our exchange in September (with Dr. Dixon, 9/10/19), you (Dr. Roberts) said:**

**“Let me elaborate and propose a compromise. It would be ideal to present a multiverse analysis that was constrained to the universe that the first paper occupied. This will be the most informative analysis for understanding whether the original authors arrived at a rare p-value in a universe of p-values related to their analytical choices. If, for example, the distribution is flat and literally only 5% of the original p-values were below .05, that would cast more doubt on the original findings than any alternative analysis you might propose.**

**The compromise would be to then propose the more appropriate analyses as you’ve done and fold them into a second multiverse analysis. “**

As proposed, we have now presented the multiverse in two parts: first, we engage in a direct replication of the logistic regression analysis used in the original paper, and second, we expand the analysis types to linear regression and propensity score analysis.

1. **“…It is probably best to eliminate any reference to Dr. Christakis himself….The goal of this paper is to correct the science, not the science communication.”**

We have removed several references to Christakis that may appear to be unnecessary criticisms, for instance the implication that he should have presented a simple scatterplot (which shows very little relation between the variables). We do not mean to imply unscrupulousness in his research techniques, and we have no intention of vilifying Dr. Christakis or any researcher who makes claims that ultimately turn out to be false. We have removed any discussion implying that his original finding was a product of p-hacking. In fact, we have removed all content related to p-hacking because we do not wish to be interpreted as making a veiled accusation.

However, although we are sympathetic to the concern that our paper will be seen as a personal attack, we do not think it’s possible to separate the science from the science communication in this instance. The original paper used careful language, yes, but the lead author publicly and repeatedly portrayed the findings as showing a causal, linear relation between TV and attention. It’s not an accident or a lack of scientific literacy that caused the media and the public to accept the claim that TV harms attention; the first author told the world that’s what the findings show. We now include an extended quote from Christakis’s 2011 TedX talk, claiming that stimulation from TV “precondition[s] the mind to expect high levels of input and … lead[s] to inattention later in life,” and that “for each hour that [children] watched before the age of three, their chances of having attentional problems was increased by about ten percent. So a child who watched two hours of TV a day before age three would be twenty percent more likely to have attention problems compared to a child who watched none (Christakis, 2011, 7:19 to 7:46), see page. 4 of our manuscript.

Similarly, we defend the decision to leave in the quote where he says “TV ‘rewires’ an infant’s brain.” Our argument is that these claims are still out there, giving validity to the argument that the 2004 paper shows a negative causal impact from TV.

We believe that scientists should be as accountable for their public statements as they are for the claims they make in scientific journals, and we do not believe that we are displaying hostility or demonstrating personal animus by quoting his verbatim statements from a high-profile public lecture.

As our goal is to show that this claim is not supported by the data, we think it is necessary to clearly identify the source of the claim in order to address it. Otherwise it may appear as if we’re creating a straw man.

1. Reviewer: 2  
     
   **Firstly, I am concerned about this manuscript’s focus on temperament and how it differentiates from attentional problems. If I would want to argue for attentional problems resulting from TV use I would say that temperament is just another measure of attentional problems and therefore conditioning on the outcome would lead to a smaller or non-significant effect. To discourage commentaries about the paper, I would suggest the authors include a theoretical discussion about why they assume that temperament is separate from attentional problems and causally prior to them in their model. If the theoretical basis for the choice to include temperament as a control is a bit unsatisfactory, I would suggest for the authors to rerun their analyses without temperament and include the results with a detailed description in the supplementary materials. Furthermore, the authors could possibly use factor analyses to probe whether temperament and attentional problems are separate constructs in an additional data-driven approach to supplement their theoretical argumentation.**

In the revised draft of the manuscript, temperament is downplayed. As you’ll see on p. 16-17, we added: “Because reviewers expressed concern that our temperament items might simply reflect attention deficits, we performed an exploratory factor analysis. A two-factor model with varimax rotation exhibited clean simple structure separating attention from temperament, and in which the largest absolute standardized cross-loading was color0.133. The correlation between factors was *r* = -0.114. We therefore concluded that attention and temperament were highly distinct variables.”

We also added a short discussion on p. 28 of how the expanded covariate set (including temperament) did not meaningfully alter the results. There is no evidence that conditioning on early temperament suppresses the association between TV and attention.  
  
**Secondly, related to the point above, I am concerned about some of the control variables that the authors decided to include. As media effects research is such a theory-poor area, I acknowledge that covariate inclusion is very difficult, but I feel like it is important to really consider closely especially as it seems to be a key part of the propensity score approach. I am especially worried about how BMI is used as a conditioning variable, as there has been ample evidence that more television viewing in younger years leads to higher BMI. As above, I would recommend the authors either include a compelling theoretical discussion for why it should be included as a control, or include re-run analyses without BMI in the supplementary materials.**

We agree with this criticism and have removed the BMI covariate from our models. We included a footnote on page 15 indicating that BMI was included in a prior version of this paper but was removed at the suggestion of a reviewer.

**Thirdly, this links in with the other two points above, I would have liked some more acknowledgements of the limitations of the authors’ approach and the conclusions that can be reached in the research area. For example, there is a clear lack of theory to build a strong theoretical argument about what variable comes first and what comes second (e.g. BMI, temperament and attentional problems). This however is important to build a causal model, and add the right covariates into it, to be able to make causal arguments. The authors try to do this using propensity score matching. I personally don’t think there is enough theory in the area to do this in a completely convincing manner, so I would also recommend the authors be more cautious about using the word “causal” in the discussion. I do not feel like our knowledge of covariates in this area is good enough for any statistical model without experimental factors to be truly causal (e.g. the use of the word ‘directional’; might make sense).**

We have edited the paper to clarify that, although we are ultimately interested in understanding the causal effect (or lack thereof) of early TV on attention, and that this is why we carefully considered covariate inclusion and the assumptions of our models, we seriously doubt that either Christakis et al’s results or ours are unbiased estimates of those causal effects (see discussion on p. 16) due to the high likelihood of residual confounding.

However, while correlation does not imply causation, causation does imply correlation. If TV and attention are not even associated, then they cannot be causally linked. In other words, an association between variables is a necessary but insufficient condition for inferring a causal relationship. Ultimately, our argument benefits from this asymmetry, as our results show that even the existence of an association between these variables is highly questionable.

**Fourth, I found parts of the multiverse argument difficult to navigate as there were many analytical choices and they were not clearly set out.** **I would recommend the authors either include a table of analytical choices (e.g. like the one included in Orben and Przybylski, 2019) or if that does not match their multiverse as it is more complicated, to include a variant on a decisions tree. This should be included early in the manuscript, for example on page 6, where it can be referenced back to throughout reading the manuscript.**

We now include Table 1 – a guide to those choices -- as the reviewer suggested. Further, while we greatly expanded the number of models we considered, the visual presentation of these results in Figures 2 and 3A-3C no longer relies on color, fill, shape, and linetype for disambiguation. We believe that the results are much more clearly presented now, and will also be much more readable in black-and-white printing as compared to the previous version of this submission. (While Figure 1 retains color, the plot features are represented and described in text such that color is unnecessary for interpretation).   
  
**Fifth, I would have enjoyed a more detailed discussion into why higher cut-offs for problematic attention lead to more significant results. I think a researcher less critical of the area than I am would find this important to interpret, so I think the paper should give that finding some space. In addition to that, I think researchers critical of this paper will zoom into how on the right panel of Figure 4 shows quite a few analyses for TV at age of 3 that look ‘significant’. This might be best addressed by adding significance testing as done in Orben and Przybylski (2019). However, I could understand if this cannot be included as I needed to keep the analyses of that paper very simple in order to be able to do the necessary simulation to do the significance testing. Whether they can implement it or not, I would like the authors to consider whether specification curve analysis-style significance testing is possible, and if not, how they can clearly argue against the existence of certain effect (this ties in with the point below).**

We have included a detailed discussion (p. 25-27) of why several of the logistic and IPTW propensity score models reached statistical significance. We propose a hypothesis and support it with four lines of reasoning, which is supported by Figures 4 and 5 as well as Tables 3 and 4. In brief, our argument is that the significant models are a result of:

* Listwise deletion, as the significance rate for multiple imputation models was much lower
* Inappropriate inclusion of sample weights in the analysis
* A subtle nonlinearity in the TV-attention scatterplot which is enhanced by logistic models with high cutpoints and IPTW models that place the dividing line between low- and high-TV groups precisely on this feature. Models overfitting this feature were far more likely to reach significance.

**Sixth, I felt like the coverage of effect sizes was quite rushed and not extensive. I would therefore recommend that the authors include some more ‘real life’ estimates of effect sizes. The manuscript would be greatly improved if the authors give a more detailed estimate of effect size, and possibly try to put that effect size into perspective. It is very difficult to communicate what a ‘small’ effect is, and a great paper would try to help the reader by giving them some additional thinking aids, e.g. what is the effect of viewing X more minutes of TV? What percentage of variance of attentional problems is explained?**

Our revision includes an expanded discussion of the implications of our effect size estimates. See p. 27.  
  
**I also have a couple of minor points:  
1. I would suggest the authors go into more detail what is meant by “it” on page 2 line 15**

We changed this to “it is crucial that scientists get the answers right.”

**2. I would add a citation for the claims made on page 2 line 45-47**

The citation is included.

**3. I would suggest the authors clarify what is meant by “during the same time frame” on page 3 lines 29-31, as all these papers were published in different years, often multiple years apart.**

It’s our goal to compare total (not cumulative) citations made across two recent years, which is a time frame after which all of the papers had been published for some time. We re-worded the sentence to hopefully clarify: “Also telling, while the original paper suggesting a link was cited 118 times in the period from January 2017 to December 2018, during the same time frame the more methodologically sound critique (Foster & Watkins, 2010) was cited 18 times and the meta-analysis (Nikkelen et al., 2014) only 38.”

**4. On page 10 the authors talk about correction for out-of-range variables and I feel like they could add some more transparency about these choices. In particular, I would recommend they say how many participants were set to missing or truncated for each factor. Also, they could explain why some were variables were set to missing, while others were truncated, the decision to do so is not clear to me.**

This is clarified in the paper. We followed the procedure described by Christakis et al. (2004), as we did not want any deviations between our results to be blamed on alterations to data cleaning and preparation. See p. 13 and 14.

**5. I would suggest the authors improve the variable naming on Table 2 and also Figure 3 so that they are more easily understandable for readers and more consistent throughout (e.g. they can make the capitalisation consistent, adhere to a more consistent use of numbers etc. e.g. In Figure 3 “Rosen87” might better be named “self-esteem”)**

The reviewer is referring to variable abbreviations in figures and tables that are no longer a part of the manuscript. These figures and tables have been moved to the extensive supporting materials on our github page under the “Results” directory. There you will find a subfolder for each analysis with its own results table, descriptive statistics, and diagnostic plots.

**6. The multiverse is also handy when the authors have had access to data, this could be added to the paragraph on page 21 lines 27-43.**

Unfortunately, we are not entirely clear on what the reviewer is suggesting here. We think she is asking us to make the point that a multiverse analysis is a good choice (an alternative to p-hacking) especially when the entire database on which the first claim was made is publicly available. We have removed the discussion of p-hacking so we do not see a place that such a comment need be inserted.   
**Once these suggestions and comments have been addressed, I would be delighted to see this paper published.**

Thank you. So would we!