**Referee Report on MS 2017412, “Quantifying the Life-Cycle Benefits of a Prototypical Early Childhood Program”**

*Reviewer #3*

The goal of the paper is to perform a comprehensive evaluation of an early childhood program (ACB/CARE). The evaluation is comprehensive in several ways.

1. It includes a wide range of outcomes.
2. It forecasts and monetizes the effects on these outcomes over the life-cycle, to estimate rates of returns and compare benefits to costs.
3. It tries to account for both the sampling error in the experimental and auxiliary samples and the forecast error due to the interpolation and extrapolation.

If I understand correctly, i) is already done in a separate paper (Garcia et al, 2017). What's new in this paper is ii) and iii), both of which are challenging tasks but potentially quite useful for program evaluation.

One way to think about the current paper is that it is a sequel to Heckman et al.'s 1997 paper 'Making The Most Out Of Program Evaluation and Social Experiments'. The 1997 paper argued (forcefully) the need to learn about the distribution of impacts (or features of it). The current paper argues it is necessary to predict and monetize long-term benefits.

As written, however, I'm afraid the current paper falls short of making a clear and compelling case for how one should combine economic theory, econometrics, and auxiliary data sets to evaluate the long-term benefits of a social experiment.

Comments:

1. For several reasons, the paper is cumbersome to read. I rarely complain about this, but:

* The paper is extraordinary long, in total 350 pages including appendices. There are numerous long footnotes. I never refereed anything like this paper. I spent several days on the manuscript and I still don't understand what exactly is being done.
* The body of the paper is far from self-contained; it is necessary to read the many appendices (and footnotes) to understand what the authors actually do, how the various procedures perform, etc. Indeed, most of the substance is in the appendices. The body of the paper actually reads a bit like a very long introduction/summary of what the authors do and what they found.

Don't get me wrong. The paper is very ambitious, and I applaud that. But after spending many hours on the paper, I'm not sure what the authors do and what I learned. According to the authors, the paper is supposed to be a template for program evaluation. If so, that template would be more useful it is was clearer and more focused.

Personally, I think it would be useful to drop the long introduction and get quickly to the core of the paper: How to forecast (and aggregate) the life-cycle costs and benefits.

Appendix C3 is important, and much of it is necessary to understand what you actually, and the pros and cons of alternative approaches.

You don't really offer a new approach to monetize the outcomes. Perhaps this part of the paper can be summarized in a short subsection with a table. The details on how you monetize given your setting may be discussed in an appendix.

1. More substantive comments:
2. Throughout the paper, the authors "test and do not reject" some testable implications of the key assumptions. Failure to reject is then taken as support of the assumptions. This raises two questions:
   * If the paper is supposed to be a template, what should the researcher do if she rejects a given assumption, such as exogeneity or structural invariance? This seems like a likely event given the strong assumptions. Is it possible to make progress under alternative or weaker assumptions?
   * In the body of the paper, the authors consistently argue that they test and do not reject the assumptions they make. Looking at the appendices, this seems to be true in many but not all cases. Also, in some of the cases it is true, the sampling error is significant and one cannot say much about the testable implications. The authors should be clear about this.
3. I don't understand how the paper deals with substitution bias. Perhaps it was discussed in one of the appendices (or in Garcia et al.) but I missed it. Please clarify what assumptions and data allow you to address this issue.
4. How do you choose predictor variables? It needs to be able to predict a treatment effect, and so the predictor variable should be affected by treatment. I might be wrong but it doesn't look like you are using the subsample of predictors that are affected by treatment? Why not? Why does it make sense to include predictor variables not affected by treatment?
5. What, if any, restrictions are there on the joint distribution of the predicted outcomes in a given period? Are, say, high earners also more likely to be healthy? What are the cross-equation restrictions in the forecasts?
6. Please clarify why one needs a randomized experiment if one is willing to make the assumptions you invoke (such as exogeneity of the predictor variables and structural invariance)? Why not just go to observational data where we can observe long-run outcomes and have a larger set of covariates to do the matching. What economic model or assumptions would invalidate such a matching procedure (which motivates why you look at social experiments), yet still be consistent with the assumptions you invoke? The paper should be clear about this.
7. The importance of essentially heterogeneity is a key insight of other work by Heckman and coauthors. Please discuss your assumptions and approaches in light of this evidence. Does the presence of essential heterogeneity in a wide range of settings tell us something of the applicability of the methods you propose in this paper?