Response to referee 2

Thank you for your comments!

- **The paper starts with a discussion of different ways that a paper can be replicated. This part of the paper is somewhat confusing because these different ways are referred to as both "steps" and "approaches." As a result, it is unclear as to whether the author is advising three different ways of doing a replication, three parts of a single approach, or some combination of the two.**

I will standardize the terminology in the revised version. Each of the steps/approaches could be done separately or one could do all three steps/approaches.

- **The author describes a successful replication for the first step, data and code from the original author, as "if the vast majority of coefficients in a paper can be replicated exactly, with typos explaining deviations" (page 1). As a guiding principle, the criteria for "successful" lacks the precision needed for different replicators to come to the same conclusion. Is the author suggesting that all the coefficients should be exactly replicated with any deviations being just typos, or is the author suggesting that only a majority of the coefficients need be exact with some difference in this majority being explained by typos? If it is the latter, how is "vast majority" being defined operationally? If it is the former, it would be much clearer to say that the replication would also be judged a success if deviations in the results can be explained as typos.**

I like this suggested formulation and will use it in the revised version

- **It is also unclear how one would judge success if all the coefficients are exact, but there are deviations in other reported numerical results, such as the significance level of coefficients, R_2, etc.**

I will replace 'coefficients' by 'numerical results'.

More general, one could compute the share of numerical results that can be replicated

- **A replication at the second step, replicator collects data and writes code, is "deemed successful even if numerical estimates are not exactly replicated but qualitatively conclusions would be confirmed" (page 1). This seems like an entirely different approach to replication rather than a second step after the exact replication described previously. If it is a second step, it is unclear why the replicator would be rewriting the code except as a check between the description of what was done in the original and what was actually done since the original code was available in the first step. The same comment can be made about the replicator collecting the data. If this is a second approach to replication, it would seem that it would be more useful to discuss the scenarios under which the replication would be deemed unsuccessful, since any deviations in the results could be due to differences in the data, the code or some combination of the two, which does not necessarily help the reader of the replication report to assess the validity of the original study. It might be very useful for the replicator to highlight differences between the data descriptions and the actual data and the description of the analysis and what was actually done if there are differences that**

significantly alter the results, but it is unclear if that is the purpose of this step or, again, if this is a different approach.

What I meant here is that one can try to replicate results based on what is described in the paper, rather than by rerunning the author's code on the author's data. In fact, this step/approach is especially useful if the code and data of the original paper is not available. If these are not available, not getting exactly the same results could be due to not all details having been clearly presented in the paper or by incorrect interpretations by the replicator. Because of this uncertainty, strong conclusions about replicability are harder to make. Though one could argue that the original code and data always should be available and absent these, any responsibility for not obtaining exactly the same results should be allocated to the original author.

If the data and code are available, this step/approach could suggest where the description in the paper by itself is insufficient to get the numerical results presented in the paper.

- **A successful replication at the third step, where the replicator tests the conclusions of the original paper with different data, is defined as whether "the general conclusion still holds" (page 1). From the description of this step at the beginning of the paper, it is not clear whether the code provided by the authors or the replicator's code is going to be used to generate the results. This is cleared up in the second part of the paper when it is indicated that the replicator uses his code, but it does raise the question of how this affects the results. It is not clear why the replication results at this step should speak to the success or failure of the original research, but rather whether the results can be generalized. If the replication results for a different time or place don't support the original conclusion, the original findings could still be useful in other circumstances. If the replicator is interested in knowing whether original conclusions apply in a particular country, such as New Zealand, then negative replication results would indicate that the original findings are not useful in New Zealand.**

I agree with the referee and will make this point clearer in the revised version. Generalization is sometimes categorized as a specific type of replication.

- **The paper would be much improved if, at the beginning, it was made clearer whether the paper**

is discussing "different ways" of replicating a paper, different "steps," or different "approaches."

I will standardize the terminology in the revised version.

- **The paper starts out by mentioning "3 steps." The first two steps are clearly delineated, but it is not totally clear on page 1 what the third step is.**

I will make this clearer in the revised version.

- **It is not clear what the purpose of the second step discussed on page 1 is. Is the replicator looking to see if there are differences between the original data and the replicator-collected data? Is the replicator looking to see if there are differences between the original code and the replicator produced code? What does the replicator do if there are significant differences? Would the replicator continue to step 3 if the replication results at step 2 don't support the original conclusions?**

See the discussion above. I will make this clearer in the revised version.