Responses to Referee Report 3

Thank you for your supportive comments and insightful suggestion.

Responses to specific suggestions are given below (following the original comments in italics).

The paper needs to make clearer the fact that, in a community of researchers that share the same paradigm, untrustworthy evidence is easy to replicate in cases where the statistical adequacy is ignored. For instance, the Efficient Market Hypothesis (EMH) and the Capital Asset Pricing Model (CAPM) has been replicated and confirmed millions of times and continue to be confirmed every day by MBA students around the world, even though a closer look at the evidence confirms that they are totally untrustworthy. This happens because the community of researchers follow the same curve-fitting procedures that give rise to very similar empirical "evidence". Hence, just because one can replicate or reproduce similar numbers and the inference results by repeating the same or similar estimation and testing procedures, does not mean that the resulting evidence are trustworthy. What makes replicability a worth-while endeavor is the emphasis on the trustworthiness of the evidence, and not on being able to get the same or very similar empirical results.

This is an excellent suggestion. I agree it is important to emphasize that the ability to reproduce the results in any study selected for replication, while worthwhile, is far from sufficient to be convinced of the trustworthiness of the results. This suggestion takes this a step further in that the results of multiple studies using similar methods and reporting similar empirical outcomes may also not be trustworthy, especially if they all neglect investigation of statistical adequacy. The EMH/CAPM example is a useful one and reference can be made to Spanos and Mayo (2015) (which will already be referenced in response to comments from Referee 1) for a nice case study of the generic approach to estimation and inference in CAPM studies. The observation that multiple studies can derive similar sorts of results because they use similar methods, but may still be untrustworthy, also has implications for meta analyses, so this will be briefly discussed in the revised version of the paper.

Spanos, A., and Mayo, D. G. (2015). Error statistical modelling and inference: Where methodology meets ontology. *Synthese*, 192, 3533–3555.