Responses to Reviewers' and Editor's Comments on "Replication to assess statistical adequacy" (Manuscript Number 2375; Discussion Paper Number 2017-73)

Editor

1) Following on the comments of Reviewer 1, it would be useful if you could place your type of replication within a taxonomy of replications. There are several taxonomies available, and you can choose whatever you think is best. I am partial to the replication types discussed in Reed (2017) (Reed, W.R., 2017, Replication in labor economics, *IZA World of Labor*, 2017:413, doi: 10.15185/izawol.413), but Clemens (2017) and Hamermesh (2007), which you already cite, would also be acceptable. To the best of my knowledge, diagnostic testing has not previously received attention within the context of replication, and finding a "place" for it among the different types of replication would give it a better foothold within the replication toolkit.

A paragraph has been added at the end of Section 2 (pp.6-7) using Reed's taxonomy. Because misspecification testing has received very little attention in replication analyses, replications testing for statistical adequacy will generally be of types 2, 4, or 6 in this taxonomy. However, footnote 5 on p.7 does discuss the conditions under which statistical adequacy replications could be of types 1, 3 or 5.

2) As you know, there has been much debate about the value of null hypothesis statistical testing. Your focus on statistical adequacy represents a full acceptance of the NHST approach. It would be illuminating to get your thoughts regarding concerns about NHST and to what extent that circumscribes the value of your approach.

The focus on statistical adequacy fits into the 'error statistical' approach, which avoids many of the excesses of NHST and provides a more coherent philosophy for statistical inference using frequentist methods (see, for example, Haig, Tests of statistical significance made sound, *Educational and Psychological Measurement*, 2017, 77(3), 489-506, for an overview). Some relevant points have been added to the discussion of misspecification testing from this perspective (on pp. 13-14), but I haven't included a review of concerns about NHST and how the error statistical approach avoids these, as this seems well beyond the scope of the paper.

3) Relatedly, and following up Reviewer 2's mention of statistical power, it would be useful if you elaborated on the interpretation of possible outcomes resulting from testing for statistical adequacy. For example, failure to reject homoskedasticity does not mean that heteroskedasticity is not a problem. It follows that failure to reject does not mean that inference is not substantially impaired. Alternatively, just because one rejects the underlying statistical assumption does not necessarily imply that the inference is substantially impaired. Any wisdom you can bring to this state of affairs would be greatly appreciated.

The revised version addresses these points, including (on p.14) a discussion of the fallacies of acceptance and rejection, linked to a wider discussion on the advantages of combining parametric and non-parametric tests and informal graphical methods to provide a 'severe' test' of the underlying statistical assumptions. The point that evidence of rejection of an underlying statistical assumption does not necessarily imply that the inferences on the primary parameters of interest in the model are substantially impaired is now clarified on p.15 and linked to the desirability of simulation evidence on the effects of different departures from the statistical assumptions.

4) Finally, be sure to respond to Reviewer 2's comment that "more tests increase the probability of rejection under the null." What would it mean if your replication conducted 20 tests for statistical adequacy, and you found that one or two of the underlying assumptions was rejected? How would you interpret that result?

Reviewer 2's point has been included in an extended discussion of several issues relevant in the choice of misspecification tests to implement (p.13). The discussion notes the desirability of choosing sufficient tests to test the relevant assumptions; however, multiple testing of different hypotheses can be taken into account by adjusting the significance level, thus controlling the overall Type I error probability. In addition, the use of joint hypothesis testing, a mix of parametric and non-parametric tests, and informal graphical analysis helps to strike a balance between probing all the relevant assumptions with adequate power while keeping the number of distinct hypothesis tests under control, so that a scenario involving conducting 20 misspecification tests can be avoided.

Referee 1

(1) On page 2, par 2, "In principle...mechanism for highlighting unreliable results in the literature." The primary purpose of replication is to confirm or deny, not just deny.

The wording has been changed to " … mechanism for distinguishing between reliable and unreliable results in the literature".

(2) Page 3, first par, bottom: "From the perspective of statistical adequacy, a replication that faithfully reproduces..." Inasmuch as on page 2 you use the phrase "reproducibility crisis", you should distinguish between reproduction and replication. While you give various definitions on page 2, you should clearly define what you mean by these terms for the purpose of your paper.

Having said the above, you should remark early on that reproduction is NOT replication, but it is where replication begins. If a paper cannot be reproduced, then there is no point in trying to replicate it.

A definition of 'reproduction' has been added, at the end of para 3 on p.2, after the statement of the Duvendack et al. (2017) broad definition of replication. A distinction is also drawn between reproduction and other forms of replication, based on the taxonomy in Reed (2018a). Given this choice of definition of replication and this taxonomy, however, reproduction is regarded as one type (the most direct type) of replication. So it is not then appropriate to classify reproduction as 'NOT replication'; this would implicitly involve an alternative definition of replication.

… You should also emphasize the role observational data play in statistical adequacy by drawing a parallel with experimental data.

For example, for an experimental paper, I will try to reproduce it. If I can, then I will try to replicate it to confirm or deny. With observational data, if I can reproduce it, I cannot replicate it because it's observational. However, I can use statistical adequacy to deny it. If it

passes statistical adequacy, while it doesn't confirm the way that a replication does for an experiment, it does lend more credence to the original study.

The status of replication with observational data depends on the definition of replication adopted. In the paper, this is interpreted in terms of Duvendack et al.'s broad definition of a replication as "any study whose main purpose is to determine the validity of one or more empirical results from a previously published study", with a specific focus on testing for statistical adequacy. Replication, defined in this way, is feasible for a study using observational data, and passing misspecification tests for statistical adequacy is a requirement for a 'successful replication'. Some brief comments are added noting that replication with experimental data and observational data have different emphases, while noting that testing for statistical adequacy also has a role to play with experimental data in ensuring that the various aspects of experimental design have been successfully applied to generate data with the expected statistical properties.

(3) page 5, line 2: "in a majority of empirical studies in economics 3." Footnote 3 does not give a citation for this claim, merely reasons for this claim. This claim needs to be documented.

This has been reworded, now citing Spanos's (2018) comment that "very few applied papers in econometric journals provide sufficient evidence for the statistical adequacy of their estimated models".

(*4) page 6, line 5: grammatically, "former colonies" is better than "ex colonies".*

This wording has been amended.

(5) page 12, last par of section 5: As I point out above, this needs to be reiterated at the beginning of the paper so that the reader can follow your argument. I already am quite familiar with spanometrics, but the reader who hasn't read Spanos will need this kind of help.

The nature of statistical adequacy is now briefly outlined in the Introduction (rather than waiting until section 2). Also, as suggested, the importance of misspecification testing as a guide to reliability of results is foreshadowed in the last paragraph of the Introduction (and reiterated in the conclusion).

Referee 2

1) There is variation in the consequences of misspecification for the properties of estimators and tests. As specification in itself is not an end goal of an empirical study, but is required to validate the quantitative results claimed, then the nuances of misspecification failure are actually relevant. For example, some forms of non-normality may lead to unknown t-statistic distributions in small samples, but could be addressed by robust inference techniques or

estimation methods that are robust to outliers. But autocorrelated disturbances in a dynamic model would lead to inconsistent parameter estimates, which is much more serious.

The revised version now includes a paragraph (in section 5) discussing the general point that some types of misspecification will have more serious consequences than others, and the implications may well be context dependent. It is also suggested that simulation analysis with artificial data structured to match the nature of the variables in the original study can provide insights into the consequences for bias and distortion of error probabilities when relevant assumptions are violated.

One suggestion would be to extend the replication study, first undertaking the relevant misspecification tests as a `pure replication' and then proceeding to find a congruent or well-specified model using the same dataset as `scientific replication' in Hamermesh's terminology. If the parameter estimates did not vary significantly then, although the initial study is invalid, its interpretation may not be. Of course, finding a vastly different congruent model would invalidate the initial study results further.

A discussion along these lines is now included on pp.15-16. Also included is an example from our recent work (Akhtaruzzaman et al., 2018) in which respecification of a model that fails tests of statistical adequacy can lead to reversal of the main conclusions in the original study.

2. What the misspecification tests aim to address is whether the model is congruent or not, i.e. does it capture the characteristics of the unknown Data Generating Process (DGP), see Bontemps and Mizon (2003). The researcher needs to define congruency in the relevant context. This is clearly done in the paper for the AGR study discussed. But it is worth emphasizing that there isn't a `one size fits all' set of misspecification tests that apply to all empirical papers. There is a trade-off, as with any statistical testing. Sufficient tests are needed to ensure congruency but they come at a price, as more tests increase the probability of rejection under the null. The tests must have the correct size properties and sufficient power when the relevant null hypothesis is false. Section 5 briefly mentions multiple testing but this is at the heart of the choice of misspecification tests.

The earlier version of the paper does not explicitly refer to ‘congruence’, but congruence and statistical adequacy are very closely related and this is now mentioned in the discussion of the LSE approach on p.5 (fn. 3). Some differences between congruence and statistical adequacy are also noted.

The brief comments on multiple testing in the earlier version have been moved to section 4 (p.13) and expanded to include the referee’s points in a broader discussion of issue relating to choice of misspecification tests. The point that there is no `one size fits all' set of misspecification tests that applies to all empirical papers is also included in this discussion.

3. The necessary tests for statistical adequacy will vary depending on the purpose of the model. If the aim is to test theory or say something about the parameters of interest, then weak exogeneity is required, and a statistically adequate specification as outlined on page 8 suffices. If the purpose is conditional forecasting, then the model also requires Granger non-causality, or strong exogeneity. And if the purpose is for policy analysis, then parameter invariance of the conditional model to interventions in the marginal model is also needed, i.e. super exogeneity. These are testable assumptions, so would fit into the misspecification framework, see Hendry (1995, ch.5).

Yes – in general, escalating degrees of exogeneity are required depending on the purpose of the analysis. Granger non-causality is testable in a well-specified VAR, but the validity of such tests would be dependent on the statistical adequacy of the VAR. However, in AGR's study, the focus is on testing theory, so weak exogeneity is the relevant concept, and the set of assumptions listed (now on p.11) is appropriate. Consequently, I haven't discussed scenarios beyond the scope of AGR's study.

4. A useful reference is Stigum (2014) in which he provides an approach to confronting theory with the data. This is similar to the Spanos approach outlined in the paper and provides motivation for the underlying argument of statistical adequacy.

This reference has been added to the discussion on p.4.

5. The paper makes clear the general principles of misspecification testing and those tests that are relevant to the AGR study discussed, focusing on relevant tests for OLS and IV (2SLS). Some comments on the relevant tests for GMM and MLE given their statistical assumptions would be helpful for readers in the general discussion.

Brief comments have been included to the effect that there is not a `one-size-fits-all' set of misspecification tests that applies to all empirical papers (p.13) and in the concluding section it is emphasized that "different estimation methods rely on different sets of probabilistic assumptions for the observed data, so the specifics of the approach discussed … will differ from other contexts". It is the case that the discussion has focused on OLS and IV estimation, the methods directly relevant for the AGR study. The nature of the data (e.g., cross-section, time-series, large-*N* small-*T* panels, moderate *N* large-*T* panels) as well as the general estimation methods (e.g., ML, GMM) would be relevant in documenting the full set of statistical assumptions imposed on the observed data and in determining a set of relevant misspecification tests. As there are many possible scenarios, fully documenting these was considered well beyond the scope of this paper.

Referee 3

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.

A discussion of how multiple studies using similar methods and reporting similar empirical outcomes may also not be trustworthy, especially if they all adopt similar methods and neglect investigation of statistical adequacy, is now included in section 5 (pp.14-15), with a reference to Spanos and Mayo's (2015) analysis of generic tests of the CAPM. The implications for meta analyses are also briefly discussed in fn. 14.