Responses to Referee Report 2

Thank you for your supportive comments and thoughtful suggestions. Responses to specific suggestions are given below (following the original comments in italics).

(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.

I agree. Some types of misspecification will have more serious consequences than others (e.g., failure of IID assumptions is likely to be more serious than non-normality), and the implications may well be context dependent. A comment to this effect will be added to the revised version of the paper. The example of autocorrelated residuals in a dynamic model (though not directly relevant to replication of AGR's cross-sectional study) can also be used as an example of the fallacy of rejection, i.e., that rejection of the null of no autocorrelation does not imply that the errors in the model are of the particular form specified under the alternative hypothesis. In this case, generalizing the dynamics in the model would be an appropriate respecification to consider (Hendry, 1995, Ch. 7).

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.

This is a very good suggestion and a comment along these lines will be included when revising the paper. Finding that a model is not statistically adequate, although informative, is not a particularly satisfying stopping point. If it is feasible to respecify an alternative, statistically adequate model, using the same data, then the results obtained can be compared with those of the original study. This is essentially the approach we used in a recent paper on the Lucas Paradox (Akhtaruzzaman et al. 2017), replicating a paper by Alfaro et al. (2008). Alfaro et al.'s resolution of the Lucas Paradox relies on inference in models that fail standard misspecification tests. Respecifying the functional form and dealing with outliers yields models that do not fail the tests, but in these models their main conclusion is reversed.

Akhtaruzzaman, M., Hajzler, C. & Owen, P.D. (2017). Does institutional quality resolve the Lucas Paradox? *Applied Economics*, doi: 10.1080/00036846.2017.1321840.

Alfaro, L., Kalemli-Ozcan, S., and Volosovych, V. (2008). Why doesn't capital flow from rich to poor countries? An empirical investigation. *Review of Economics and Statistics* 90, 347-368.

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.

I agree. The current version of the paper does not explicitly refer to 'congruence', but congruence and statistical adequacy are very closely related and this point will be added in the brief discussion of the LSE approach on p.4. However, it is also worth noting that there are some differences: congruence involves a mixture of statistical and substantive criteria, whereas statistical adequacy is defined purely in terms of statistical assumptions (Spanos, 2006, p.41), which makes the distinction between statistical and structural/theory models 'cleaner'.

I also completely agree that there isn't a `one size fits all' set of misspecification tests that apply to all empirical studies. The set of relevant misspecification tests will vary depending on the estimation methods used and the nature of the data. In addition, the Type I error of the overall set of misspecification tests needs to be kept under control. The brief comment on multiple testing will be expanded to make these points and to mention the usefulness of nonparametric misspecification tests (Spanos,2017) (as well as joint testing already briefly referred to on p.11).

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, increasingly stronger notions of exogeneity are required depending on the purpose of the analysis. In AGR's study, the focus is on testing theory and weak exogeneity is the relevant concept. A comment will be added to make this explicit. It is also worth noting that testing for exogeneity (e.g., using a Durbin-Wu-Hausman-type test) requires the

multivariate linear regression model, made up of the reduced forms, to be statistically adequate and for any overidentification restrictions not to be rejected (Spanos, 2007). In other contexts, super exogeneity could be tested, for example, using impulse-indicator saturation in the marginal models of the assumed exogenous variables, and then adding the significant indicators to the conditional model and testing their significance (Hendry and Doornik, 2014, Ch.22).

Hendry, D. F., and Doornik. J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Model Selection Methods in Econometrics*. Cambridge, MA: MIT Press.

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.

Thank you for this reference. This could be added to the discussion, although Stigum's overall approach is less accessible than that of Spanos or Hendry.

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.

Apart from the brief comment in the concluding section 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 solely on OLS and IV estimation, the methods directly relevant for the AGR study. As noted above, 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, documenting these was considered beyond the scope of this paper, but this general point could be made more forcefully.