January 20, 2018
Dear Professor Owen,

I have now received reports from three reviewers regarding your manuscript, "Replication to assess statistical adequacy," (*Manuscript Number 2375; Discussion Paper Number 2017-73*).

Based on the feedback I have received, I am inviting you to submit a revision. Your revision should address the individual points raised by each of the reviewers. In addition, you should include a separate document that provides a point-by-point response to each of the reviewer's comments. You are not required to do everything the reviewers say, but if you disagree with a recommendation, you should state your reason. A successful revision will include addressing the following:

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.

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.

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.

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?

A revision that satisfactorily addresses the points above, as well as responding to the individual reviewers' concerns, is likely to result in a favourable publication decision.

Thank you for submitting your manuscript to *Economics: The Open-Access, Open-Assessment E-Journal*. I look forward to receiving your revision.

Sincerely,
W. Robert Reed
Co-Editor, *Economics E-Journal*