January 20, 2018

Dear Professor McCullough,

I have now received reports from three reviewers regarding your manuscript, "Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive," (*Manuscript Number 2356; Discussion Paper Number 2017-78*).

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) Reviewer 1 says that the replication you propose would not lead to any useful insights, because the original Glandon (2011) study was not well-designed. I agree. Therefore, you should expand your replication plan to consist of two parts. The first part would be the replication of Glandon (2011) that you currently propose, which is largely fine as is. The second part would consist a replication plan that described how you think Glandon (2011) should have been done. The latter would be a different "type" of replication.

In terms of the taxonomy below (Reed, W.R., 2017, Replication in labor economics, *IZA World of Labor*, 2017:413, doi: 10.15185/izawol.413), the replication you propose is a Type 1 replication ("A Reproduction"). However, as is evident in the other contributions to the Special Issue, there are other types of replications. The replication that Reviewer 1 suggests you do is a Type 4 replication ("Robustness analysis – same population").

Figure 1. Six different kinds of replications

Measurement and/or analysis	Source of data/population		
	Same data set	Same population	Different population
Same	(1) Reproduction	(3) Repetition	(5) Extension
Different	(2) Robustness analysis —same data set	(4) Robustness analysis —same population	(6) Robustness analysis —different population

Source: Author's own compilation.

IZA World of Labor

In addition to describing how you would carry out this latter replication, you should also explain how you would compare the results from the second replication with those from Glandon (2011) and the first replication, and how you would interpret that comparison. For example, in light of your stated purpose of determining whether "all was well with the AER's archive", what kind of results would lead you to conclude that "all was well." Would anything less than 100% reproducibility be grounds to conclude that all was not well? Some fleshing out of how you would interpret the results from this extended replication would be interesting and insightful.

2) Reviewer 2 notes that the paper reads like investigative reporting, and inappropriately emphasizes the role of Robert Moffit and Philip Glandon. I agree. There is no need to browbeat Professor Moffit and Philip Glandon. The purpose of the Special Issue is to further an understanding of how replications should be done. Not to assign blame for past, failed reproducibility policies. The paper should be rewritten to minimize the discussion of the roles of Moffitt and Glandon, so that the focus can properly be placed on how best to do the replication.

3) Reviewer 3, and your subsequent response, suggest some ambiguity about the role of a data archive. Is the purpose of the data archive to allow a push-button reproduction of the results in the original paper? So that given the author's final version of the dataset and the accompanying code, hitting the "button" and running the programs would reproduce the original results? Or does your understanding of a data archive require that authors give a complete and thorough description, with accompanying code, which allows one to reproduce the results from the primary data sources? Some discussion of the pros and cons of each of these would be helpful, as currently you have left vague the expectations that should be placed on a data archive. This is necessary to determine whether "all is well" with the data archive. Note that Reviewer 3's comment about Hoxby's example implicitly refers to the issue of what a data archive should be expected to do.

4) Finally, as noted by Reviewer 2, and as acknowledged in your response, you should provide some discussion about how to handle numerical differences across software and hardware applications when making "binary" decisions about whether an original study is "reproducible."

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*