Owen, Dorian

Working Paper

Replication to assess statistical adequacy

Economics Discussion Papers, No. 2017-73

Provided in Cooperation with:
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Owen, Dorian (2017) : Replication to assess statistical adequacy, Economics Discussion Papers, No. 2017-73, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/169132

Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2017-73 | September 28, 2017 | http://www.economics-ejournal.org/economics/discussionpapers/2017-73

Replication to assess statistical adequacy

Dorian Owen

Abstract
'Statistical adequacy' is an important prerequisite for securing reliable inference in empirical modelling. This paper argues for more emphasis on replication that specifically assesses whether the results reported in empirical studies are based on statistically adequate models, i.e., models with valid underpinning statistical assumptions that satisfy relevant diagnostic tests of misspecification. A replication plan is briefly outlined to illustrate what this would involve in practice in the context of a specific study by Acemoglu, Gallego and Robinson (Institutions, human capital, and development, *Annual Review of Economics*, 2014).

(Published in Special Issue The practice of replication)

JEL C31 C36 I25 P14 O10
Keywords replication; statistical adequacy; inference; instrumental variables; reduced form; fundamental determinants of economic development

Authors
Dorian Owen, ✉ University of Otago, Dunedin, New Zealand,
Dorian.Owen@otago.ac.nz

Citation Dorian Owen (2017). Replication to assess statistical adequacy. Economics Discussion Papers, No 2017-73, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-73

Received September 13, 2017 **Accepted as Economics Discussion Paper** September 25, 2017
Published September 28, 2017

1. Introduction

Widespread difficulties in replicating scientific results, whether from observational or experimental studies, have received considerable recent attention (e.g., National Academies of Sciences, Engineering, and Medicine, 2016). Concerns over the inability to reproduce results in previous published studies have been characterized as a "reproducibility crisis" affecting multiple disciplines in the sciences, biomedicine and the social sciences (Ioannidis and Panagiotou, 2011; Begley and Ellis, 2012; Doyen et al., 2012; Button et al., 2013; Open Science Collaboration, 2015; Dumas-Mallet et al., 2016; Baker, 2016).

In principle, replication of existing studies provides a mechanism for highlighting unreliable results in the literature. Conventionally, however, replication has not been a favoured activity for a variety of reasons (Duvendack Palmer-Jones and Reed, 2017), including the pressure to publish, in a culture that rates novelty more highly than accuracy.[1] Moreover, there is no consensus on what constitutes a 'replication' and different criteria and guidelines have been proposed (e.g., Hamermesh, 2007; National Academies of Sciences, Engineering, and Medicine, 2016; Hubbard, 2016; Clemens, 2017). Pragmatically, Duvendack et al. (2017, p.47) define a 'replication' broadly as "any study whose main purpose is to determine the validity of one or more empirical results from a previously published study".

The aim of the current paper is to argue for an approach to replication that specifically assesses whether the results reported in empirical studies, especially those using observational data, are based on 'statistically adequate' models, and to briefly outline a replication plan to illustrate what this would involve in practice in the context of a specific study. This fits into the broad definition of replication proposed by Duvendack et al. (2017), but is much more sharply focused on testing for misspecification of the underlying

[1] The Economist (2013), for example, cites psychologist Brian Nosek's comment that "[t]here is no cost to getting things wrong … The cost is not getting them published".

probabilistic assumptions in published studies. In contrast, existing replication studies are usually more concerned with reproducing results from previous studies or seeing if they extend to new data, different estimation methods or variations in model specifications.

In section 2, it is argued that if assessment of the reliability of inferences in empirical studies is the goal of replication, then it is important to examine the extent to which the probabilistic assumptions of the methods and models used are appropriate for the data at hand. The motivation for selecting the study by Acemoglu, Gallego and Robinson (AGR) (2014) for replication, as an illustration of what this would involve, is discussed in section 3. The type of data and estimation methods used in AGR's study affect the details of the approach to testing statistical adequacy in the replication plan summarized in section 4. Different estimation methods applied to different types of data would affect the specifics of implementation, but the underlying aim of assessing the validity of the probabilistic assumptions underpinning estimation and inference would be a common theme in the proposed focus for such replication exercises. A summary of what constitutes a 'successful' replication is contained in section 5. From the perspective of statistical adequacy, a replication that faithfully reproduces results from the original study or assesses how different the results become with different data or specifications would not necessarily provide the required insights to judge whether estimation and inference in the original study is reliable. Section 6 contains some brief concluding comments.

2. Replication to assess statistical adequacy

Economics is a discipline that relies heavily on empirical evidence, but econometric estimation and testing often appears to focus on quantifying the 'presumed-true' economic theory model (i.e., obtaining estimates and establishing statistical significance of key parameters). This form of empirical analysis becomes essentially a 'curve fitting' exercise

(Spanos, 2015). The end result of such an approach is to 'illustrate' the theory, rather than rigorously test tentative economic theory conjectures against the data (Gilbert, 1986).

In the context of empirical modelling in economics, it is helpful to distinguish between the *theory model*, which contains the substantive content based on economic theory, and the *statistical model* that is taken to the data (Spanos, 2015).[2] The statistical model (as opposed to the substantive economic theory content of the model) is the complete set of probabilistic/statistical assumptions imposed on the data. These probabilistic assumptions vary depending on which econometric or statistical technique is applied to the data. For example, in the conventional multiple regression model the assumptions include normality, linearity, homoskedasticity, independence, and constant parameters (e.g., Spanos, 2017, Table 9). A statistical model is considered to be 'statistically adequate' when all its probabilistic assumptions are valid for the observed data (Spanos 2017). The appropriateness of these underpinning statistical assumptions is crucial for securing reliable inference (Spanos, 2015, 2017). If the statistical assumptions are invalid for the data to which the statistical model is being applied, then the sampling distributions of the test statistics that are being used for inference will not be appropriate and nominal significance levels will be misleading. The end result is unreliable inference.

Misspecification (diagnostic) testing plays a crucial role in probing whether the probabilistic assumptions of whatever statistical technique is being used are valid for the data under consideration and, as a result, in securing trustworthy inference (Spanos, 2017). This view is not new and has long been a feature of the 'LSE approach' to econometric modelling (Hoover, 2006; Hendry, 2009). As McAleer (1994, p.329) notes, "[a]lthough there are dissenters, a consensus seems to have developed among sensible data analysts that diagnostic tests are essential in evaluating econometric models". However, such testing appears to have

[2] In contrast, these features of the empirical model are usually rolled into one, typically by attaching a stochastic error term, which is assumed to satisfy a set of statistical properties, to an economic-theory-based structural model.

become less common in recent years, with very limited or quite often no diagnostics reported in a majority of empirical studies in economics.[3] A more common response to uncertainty about the specification of empirical models is to conduct a robustness analysis by adding control variables, either in sets or one at a time, to regressions that include the key explanatory variable(s) of interest. However, without explicit misspecification testing, there is no guarantee that all, or indeed any, of these models are statistically adequate.

Given that "[s]cience is about inference" (King, 2017), assessment of the reliability of empirical results is the primary motivation for replication. The dependence of reliable inference on the appropriateness of the underlying probabilistic assumptions imposed on statistical models opens up an important role for replication analyses in probing the statistical adequacy of existing studies through the application of misspecification testing of the full set of such assumptions.

3. The candidate study selected for replication

The candidate paper selected for replication is a study by Acemoglu, Gallego and Robinson (AGR) (2014). This is a recent high-profile contribution to a thriving literature on the fundamental determinants of economic development.[4] Rather than explaining long-run growth and development based on 'proximate' determinants of growth (such as physical capital accumulation and technological progress), this literature focuses on 'deeper', more fundamental, determinants of levels of economic development, such as geography, institutions, history, biology and culture. An early, highly influential, study by Acemoglu, Johnson and Robinson (2001) emphasizes the quality of institutions as the key determinant of long-run economic development. It introduces (the logarithm of) historical European settler

[3] Some of the reasons for this are discussed by Spanos (2017, section 4), who also provides a robust and detailed critique of claims that discourage misspecification testing.
[4] Although published relatively recently, Acemoglu et al.'s (2014) paper has already accrued 145 Google Scholar citations (as at 12 September 2017).

mortality rates as an instrument for current institutions, to allow for the latter's endogeneity arising from reverse causality, omitted variables, and measurement error. Estimates of the effect of institutional quality, proxied by a measure of the strength of property rights, on the log of GDP per capita in 1995 are quantitatively large and statistically significant for their sample of ex colonies. However, Glaeser et al. (2004) challenge this interpretation and argue that, rather than institutions, it was the human capital brought by settlers to their colonies that had a greater effect on current levels of development.

AGR address this difference in views by including both institutional quality and human capital measures in cross-country regressions explaining real GDP per capita in 2005. As both institutions and human capital are plausibly endogenous explanatory variables, both require instrumenting. AGR follow their earlier studies in using settler mortality (capped at a maximum level of 250 per 1,000 people per annum), as in AJR (2012), and the log of population density as the main instruments for institutions (proxied by the Worldwide Governance Indicators' Rule of Law index (Kaufmann, Kraay and Mastruzzi, 2013). For human capital (proxied by average years of schooling), they use the number of Protestant missionaries per 10,000 people in the 1920s, following Woodberry (2012), and primary school enrolment rates (relative to the population aged 6 to 14) in 1900 as additional instruments. Different sets of control variables are included in the various models considered, including latitude, continental dummies, and dummies for British and French colonies. Results are reported for ordinary least squares (OLS), two-stage least squares (2SLS) and limited information maximum likelihood (LIML) estimation, and also for semi-structural models in which either institutional quality or human capital is instrumented while the instruments for the other endogenous explanatory variable are directly included. Their results strongly support the view that institutional quality is the key fundamental determinant of long-run development, in line with the conclusions of Acemoglu et al. (2001), whereas the

effects of human capital are quantitatively roughly in line with micro estimates of the return to schooling but are generally not statistically significant.

This study is an interesting candidate for replication because it provides a sharp conclusion on the institutions versus human capital debate, an important point of contention in the literature, in a framework that explicitly addresses endogeneity of both key variables. Data sources and methods are clearly summarized in the paper. Data and Stata code are available at https://economics.mit.edu/faculty/acemoglu/data/hcapital, so there are unlikely to be problems in reproducing the results reported in the paper.[5] This allows the replication analysis to focus attention on testing for statistical adequacy.

Replication of this study provides a natural extension to earlier work reported by Owen (2017), which implements misspecification testing of the reduced forms (RFs) associated with instrumental variables (IV) estimation in selected influential studies in the literature on the fundamental determinants of economic development.[6] This testing reveals widespread evidence of model misspecification, with parameter non-constancy and spatial dependence of the residuals being a widespread problem. This potentially undermines the inferences drawn about the structural parameters being estimated in these studies. Although AGR's study addresses the endogeneity of both institutions and human capital, it shares several characteristics of the earlier studies that revealed evidence of misspecification; these include the highly parsimonious nature of the structural models, lack of testing of underlying statistical assumptions, relatively modest sample sizes as a basis for relying on asymptotic results (N = 62 for the cross-country estimates), and evaluation of robustness of results by adding a relatively limited set of control variables, either singly or in sets. The diagnostic

[5] Comments in the Stata do files point out that the available data set includes a correction for Hong Kong that will lead to minor differences in some of the results reported in the paper.

[6] Owen (2017) considers misspecification testing of RFs corresponding to selected IV estimates from the studies by Hall and Jones (1999), Acemoglu et al. (2001), Easterly and Levine (2003), Sachs (2003), Ashraf and Galor (2011), and Ashraf and Galor (2013). Illustrative models from the studies by Spolaore and Wacziarg (2009), Putterman and Weil (2010), and Easterly and Levine (2016), reported by Spolaore and Wacziarg (2013) in their review article, are also examined.

tests that AGR report are limited to tests of underidentification (Kleibergen and Paap, 2006), overidentifying restrictions (Hansen, 1982), and *F*-tests on the coefficients of the excluded instruments in the first-stage regressions. However, as Spanos (2007) emphasizes, the validity of these tests is conditional on the statistical adequacy of the RFs.

For all the studies examined by Owen (2017), the country is the unit of geographical aggregation, so estimation relies on cross-country variation in the variables. AGR also consider cross-regional variation from 684 regions from 48 countries, although due to lack of data on institutional quality the models fitted to the regional data focus on the effects of human capital on development.[7] One interesting question that can be addressed with AGR's regional data is whether the evidence of spatially correlated residuals evident in most of the country-level studies is also present in sub-national data.

4. Replication plan

Testing for statistical adequacy involves testing the full set of probabilistic assumptions underpinning estimation and inference in the specific application at hand. In the case of AGR's study, the estimation methods used include 2SLS and LIML, which address the endogeneity of institutions and human capital. In this context, the replication follows the approach proposed by Spanos (1990, 2006, 2007, 2015), and applied in the context of selected studies of the fundamental determinants of development by Owen (2017). Spanos's overarching argument is that "theory-based concepts like structural parameters, structural errors, orthogonality and non-orthogonality conditions, gain statistical 'operational meaning' when embedded into a statistical model specified exclusively in terms of the joint distribution of the *observable* random variables involved" (Spanos, 2007, p.39, emphasis in original). In IV estimation, the relevant statistical model specified in terms of the observable variables is

[7] Human capital is again proxied by average years of schooling, and instrumented by a dummy for the presence of a Protestant mission station in the region in 1916.

the multivariate linear regression model consisting of the full set of RFs (including the RF for the dependent variable as well as the endogenous explanatory variables), which depends on the specification of the structural model and the associated instrumentation strategy. The multivariate linear regression model made up of the RFs provides a framework in which the structural model is embedded. A key insight of Spanos's analysis is that assumptions about endogeneity of some of the explanatory variables and exogeneity of the instruments (which are not directly testable because of the unobservable nature of the error term in the structural model) are 'operationalized' via the reparameterization/restrictions implied on the statistical model, i.e., the set of RFs. Because the structural model is a reparameterized/restricted version of the RFs, "the statistical adequacy of the latter ensures the reliability of inference in the context of the former" (Spanos 2007, p. 48). This approach is discussed in detail by Spanos (2007) and summarized by Owen (2017, section 3).

Inference, based on conventional formulae, will be appropriate if the following probabilistic assumptions apply to the multivariate linear regression model, made up of the RFs (Spanos, 2007, Table 2.2):

Normality	$D(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i}, \boldsymbol{\Theta})$ is normally distributed	(1)
Linearity	$E(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i})$ is linear in $\boldsymbol{Z}_i$ and $\boldsymbol{X}_{2i}$	(2)
Homoskedasticity	$Var(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i}) = \boldsymbol{\Omega}$ is homoskedastic (free of $\boldsymbol{Z}_i$, $\boldsymbol{X}_{2i}$)	(3)
Independence	$(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i})$, $i = 1, 2, \ldots, N$ are independent random variables	(4)
i-invariance	$\boldsymbol{\Theta}$ is constant for all i	(5)

D(.) denotes the joint distribution, and $\boldsymbol{y_i} = (y_i, \boldsymbol{X}_{1i}')'$, where y is the dependent variable in the structural equation of interest, $\boldsymbol{X}_{1i}$ is a vector of endogenous explanatory variables, $\boldsymbol{X}_{2i}$ a vector of exogenous explanatory variables, $\boldsymbol{Z}_i$, a vector of additional instruments that satisfy exclusion restrictions, and $\boldsymbol{\Omega}$ is the error covariance matrix and $\boldsymbol{\Theta}$ a vector of parameters in the multivariate linear regression. Subscript i denotes observations for country i ($i = 1, \ldots, N$).

Assessment of statistical adequacy of the multivariate linear regression model made up of the RFs involves testing these assumptions. This approach contrasts sharply with common practice in applications of IV estimation, which ignores the embedding nature of the set of RFs and treats fitting a linear projection in first-stage regressions as purely a predictive exercise. It is also common to appeal to a weaker set of assumptions to justify the asymptotic properties of 2SLS estimation and to use asymptotically valid heteroskedastic-robust standard errors for inference. However, Owen (2017, p.8) argues that, especially for the modest sample sizes typically found in the fundamental determinants literature (here $N = 62$), reliance on asymptotic results that depend on a weaker set of implicit and untested (or untestable) assumptions is less appealing than basing inference on a statistical framework subject to a set of explicit non-rejected assumptions.[8]

The assumptions in (1)-(3) can be tested using conventional tests for normality (Doornik and Hansen, 2008), functional form (Ramsey's (1969) RESET test) and heteroskedasticity (White, 1980). Given the MLR nature of the RFs, system misspecification tests, multivariate equivalents of these single-equation tests (e.g., Doornik and Hendry, 2013, p. 227), can also be examined. With cross-country data, failure of the independence assumption in (4) is likely to involve spatial dependence, interpreted broadly to include dependence based on socio-economic as well as geographical distance. Spatial dependence can be tested using Moran's I statistic (Moran, 1948) and/or a Lagrange Multiplier (LM) test (Anselin et al., 1996) applied to the residuals of the fitted RFs, with the required a priori weights matrix based on plausible assumptions about the extent of potential spatial linkages.

Parameter constancy in (5) can be examined by recursive graphical analysis of coefficient estimates for the variables in the RF and also of break-point Chow tests at different points in the sample (Hendry and Nielsen, 2007, pp.195-197). Different orderings of cross-sectional

[8] See also Spanos (2015, p.183; 2017, section 4.4.4) on the disadvantages of methods that rely on weaker assumptions for their asymptotic properties.

will affect the recursive plots and Chow tests, but ordering the observations by the log of GDP per capita revealed patterns of interest in the studies examined by Owen (2017), so this would be a natural choice.[9]

In addition to, or as an alternative to, testing the different statistical assumptions using separate tests, Spanos (2017) recommends joint testing using auxiliary regressions that incorporate terms to allow for departures from the various assumptions.

If the RFs appear to be statistically adequate, it is then appropriate to test for weak instrumentation (e.g., using Cragg and Donald's (1993) test in conjunction with Stock and Yogo's (2005) critical values) and overidentifying restrictions (Sargan, 1958; Hansen, 1982) as their validity is conditional on the statistical adequacy of the RFs (Spanos, 2007).

5. What constitutes a successful replication?

If, as suggested, the focus of a replication exercise is on the statistical adequacy of the models on which inference is based, then a successful replication would reveal little or no evidence of misspecification (i.e. failure of the underlying probabilistic assumptions) for a set of results that reproduces those reported in the original study. Such an evaluation would need to take into account multiple testing of different hypotheses, for example by selecting a numerically smaller significance level (e.g., 1% instead of 5%) for each test. Also, misspecification test results can be considered holistically, as rejection of a specific null hypothesis may not provide a clear guide to the type of misspecification (e.g., lack of parameter constancy could arise for a number of reasons, including outliers, omitted variables, heteroskedasticity, etc.).

If the RFs are found to be statistically adequate, and subsequent testing does not reject overidentifying restrictions or raise concerns about weak instrumentation, then inference on the structural parameters of interest, such as the coefficients on institutions and human capital

[9] The various tests and their interpretation are discussed in more detail in Owen (2017, Section 4).

in the models for the level of economic development, can proceed and the substantive economic theory contribution of the models evaluated. At this point, provided the point estimates, standard errors and other reported statistics in the original study are reproducible (as seems likely in the case of AGR's study), then there would be no reason to call into question the reliability of inference on the structural parameters.

From this perspective, the ability to take the original data set and exactly reproduce the reported results of the original study would represent a necessary but far from sufficient condition to constitute a 'confirmation' of the results. If significant evidence of misspecification were to be found, this would point to a 'disconfirmation', or at least flag the need for additional analysis.

6. Concluding comments

The primary motivation of this paper is to make a case for more emphasis on testing for statistical adequacy in replication analyses. If we are to trust the results in the empirical literature in economics, we need to verify the statistical underpinnings of the various models that we estimate and use as a basis for inference. Different estimation methods rely on different sets of probabilistic assumptions for the observed data, so the specifics of the approach discussed above for the RFs for IV estimation (which are at odds with common practice) will differ from other contexts. However, a common feature of the approach would be an emphasis on misspecification testing of the full set of probabilistic assumptions imposed on the observed data.

References

Acemoglu, D., Gallego, F. A., and Robinson, J. A. (2014). Institutions, human capital, and development. *Annual Review of Economics*, 6, 875-912.

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91, 1369-1401.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2012). The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, 102, 3077-3110.

Anselin, L., Bera, A. K., Florax, R., and Yoon M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26, 77–104.

Ashraf, Q., and Galor, O. (2011). Dynamics and stagnation in the Malthusian epoch. *American Economic Review*, 101, 2003–2041.

Ashraf, Q., and Galor, O. (2013). The "out of Africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 103, 1–46.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454.

Begley, C. G., and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483, 531–533.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365–376.

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 31, 326–342.

Cragg, J. G., and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9, 222–240.

Doornik, J. A., and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70, 927–939.

Doornik, J. A., and Hendry, D. F. (2013). *Modelling Dynamic Systems, PcGive 14, Volume II*. London: Timberlake Consultants Ltd.

Doyen, S., Klein, O., Pichon, C. L., and Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7, e29081.

Dumas-Mallet, E., Button, K., Boraud, T., Munafo, M., and Gonon, F. (2016). Replication validity of initial association studies: A comparison between psychiatry, neurology and four somatic diseases. *PLoS ONE*, 11, e0158064.

Duvendack, M., Palmer-Jones, R., and Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review: Papers & Proceedings*, 107, 46-51.

Easterly, W., and Levine, R. (2003). Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics*, 50, 3–39.

Easterly, W., and Levine, R. (2016). The European origins of economic development. *Journal of Economic Growth*, 21, 225–57.

The Economist (2013). Unreliable research: Trouble at the lab. *The Economist*, Oct 21st. Available online: https://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble (accessed on 3 May 2017).

Gilbert, C. L. (1986). Professor Hendry's econometric methodology, *Oxford Bulletin of Economics and Statistics*, 48, 283-307.

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., and Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, 9, 271-303.

Hall, R. E., and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics*, 114: 83–116.

Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics*, 40, 715-733.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking glass. In T.C. Mills and K. Patterson (Eds), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*. Basingstoke: Palgrave Macmillan, pp. 3-67.

Hendry, D. F., and Nielsen, B. (2007*). Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University Press.

Hoover, K. D. (2006). The methodology of econometrics. In T. C. Mills and K. Patterson (Eds), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*. Basingstoke: Palgrave MacMillan, pp. 61-87.

Hubbard, R. (2016). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. Thousand Oaks, CA: Sage Publications.

Ioannidis, J. P. A., and Panagiotou, O. A. (2011). Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *Journal of the American Medical Association*, 305, 2200–2210.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2013). Worldwide Governance Indicators (www.govindicators.org), The World Bank.

King, G. (2017). Gary King discusses replication in the social sciences. Sage Research Methods Video, http://dx.doi.org/10.4135/9781473999916 (accessed on 2 September 2017).

Kleibergen, F., and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133, 97-126.

McAleer, M. (1994). Sherlock Holmes and the search for truth: a diagnostic tale. *Journal of Economic Surveys*, 8, 317-370.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10, 243–251.

National Academies of Sciences, Engineering, and Medicine (2016). *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*. Washington, DC: The National Academies Press.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716.

Owen, P. D. (2017). Evaluating ingenious instruments for fundamental determinants of long-run economic growth and development. *Econometrics*, 5, 38; doi:10.3390/econometrics5030038, http://www.mdpi.com/2225-1146/5/3/38/pdf

Putterman, L., and Weil, D. N. (2010). Post-1500 population flows and the long-run determinants of economic growth and inequality. *Quarterly Journal of Economics*, 125, 1627–1682.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B*, 31, 350–371.

Sachs, J. D. (2003). Institutions don't rule: Direct effects of geography on per capita income. NBER Working Paper 9490, National Bureau of Economic Research, Cambridge, MA.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26, 393–415.

Spanos, A. (1990). The simultaneous-equations model revisited: Statistical adequacy and identification. *Journal of Econometrics*, 44, 87–105.

Spanos, A. (2006). Econometrics in retrospect and prospect. In T. C. Mills and K. Patterson (Eds), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*. Basingstoke: Palgrave MacMillan, pp. 3–58.

Spanos, A. (2007). The instrumental variables method revisited: On the nature and choice of optimal instruments. In G. D. A. Phillips and E. Tzavalis (Eds), *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*. Cambridge: Cambridge University Press, pp. 34–59.

Spanos, A. (2015). Revisiting Haavelmo's structural econometrics: Bridging the gap between theory and data. *Journal of Economic Methodology*, 22, 171–196.

Spanos, A. (2017). Mis-specification testing in retrospect. *Journal of Economic Surveys*, doi: 10.1111/joes.12200.

Spolaore, E., and Wacziarg, R. (2009). The diffusion of development. *Quarterly Journal of Economics*, 124, 469–529.

Spolaore, E., and Wacziarg, R. (2013). How deep are the roots of economic development? *Journal of Economic Literature*, 51, 325–369.

Stock, J. H., and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews and J. H. Stock (Eds), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

Woodberry, R. D. (2012). The missionary roots of liberal democracy. *American Political Science Review*, 106, 244-274.