Economics
The Open-Access, Open-Assessment E-Journal
Vol. 12, 2018-60 | September 21, 2018 | http://dx.doi.org/10.5018/economics-ejournal.ja.2018-60

Replication to assess statistical adequacy

P. 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 pass relevant diagnostic tests for 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; reliability of inference; science trustworthyness; misspecification testing; instrumental variables; reduced form; fundamental determinants of economic development

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

Citation P. Dorian Owen (2018). Replication to assess statistical adequacy. *Economics: The Open-Access, Open-Assessment E-Journal*, 12 (2018-60): 1–16.
http://dx.doi.org/10.5018/economics-ejournal.ja.2018-60

Received September 13, 2017 **Published as Economics Discussion Paper** September 28, 2017
Revised September 10, 2018 **Accepted** September 17, 2018 **Published** September 21, 2018

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, 2005; 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; Hubbard, 2016).

In principle, replication of existing studies provides a mechanism for distinguishing between reliable and 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 emphasizes quantity to the detriment of quality (Smaldino and McElreath, 2016; Edwards and Roy, 2017) and rates novelty more highly than replicability (Grimes, Bauch and Ioannidis, 2017; The Economist, 2017).[1]

Moreover, there is no consensus on what exactly 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; Galiani, Gertler and Romero, 2017; Reed, 2018a, 2018b). Pragmatically, Duvendack et al. (2017: 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". This could involve anything from 'reproduction' of results (i.e., using exactly the same variable definitions, data and methods as in the original study), through to 'robustness analyses' that employ different measures, different methods and/or data from different populations (e.g., Reed, 2018a, Table 1).

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. Statistical adequacy refers to the validity of the probabilistic assumptions imposed on the stochastic process underlying the data (Spanos, 2015). Probing statistical adequacy fits into the broad definition of replication proposed by Duvendack et al. (2017), but is more sharply focused on testing for misspecification of the underlying probabilistic assumptions in published studies. To date, replication studies have neglected this aspect of study reliability; they 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 specification.

In Section 2 it is argued that, if assessment of the reliability of inferences in empirical studies is the goal of replication, 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 determine the specifics of testing for

[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".

statistical adequacy in the replication plan summarized in Section 4. Different estimation methods applied to different types of data would affect the details 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 discussion of what constitutes a 'successful' replication is contained in Section 5. From the perspective of statistical adequacy, the underlying probabilistic assumptions would need to be probed using misspecification tests; detection of significant departures from these assumptions would call into question the reliability of the primary results, even if these results appear to accord (e.g., in terms of statistically and quantitatively significant effects) with those in the original study. If there is no attempt to probe the probabilistic assumptions underlying inference, then a faithful reproduction of the results from the original study, or a replication that produces similar results with different data or specifications, would not provide the required insights to judge whether estimation and inference in the original study are 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 appear 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 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, 2006, 2015; Hendry, 2009, 2015: Ch.4; Stigum, 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, 2018: Table 9). A statistical model is considered to be 'statistically adequate' when all its probabilistic assumptions are valid for the observed data (Spanos, 2018). The appropriateness of these underpinning statistical assumptions is crucial for securing reliable inference (Spanos, 2015, 2018). 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 error probabilities will be potentially misleading. The end result is unreliable inference.

[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 model.

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, 2018). This view is not new and has long been a feature of the 'LSE approach' to econometric modelling (Hoover, 2006; Hendry, 1995, 2009).[3] McAleer (1994: 329) notes that "[a]lthough there are dissenters, a consensus seems to have developed among sensible data analysts that diagnostic tests are essential in evaluating econometric models".

Unfortunately, "very few applied papers in econometric journals provide sufficient evidence for the statistical adequacy of their estimated models" (Spanos, 2018: 555).[4] 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. More recently, the 'design-based' approach to microeconometrics (e.g., Angrist and Pischke, 2010) has emphasized research design as the key to identifying causal effects, while downplaying traditional econometric concerns such as misspecification. In contrast, Nevo and Whinston (2010: 80) argue that empirical work (in micro and macroeconometrics) needs to combine "careful design, credible inference, robust estimation methods, and thoughtful modelling … this should not be an either-or proposition".

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. Empirical studies in economics are usually based on observational data, but laboratory experiments and randomized control trials (RCTs) are becoming increasingly common. While there are differences in emphasis in replicating results from experimental data (Spanos, 2010; Camerer et al., 2016), testing for statistical adequacy also has a role to play in ensuring that the various aspects of experimental design have been successfully applied to generate data with the expected statistical properties (Spanos, 2010; Spanos and Mayo, 2015). In a similar vein, Brown and Wood (2018), in this Special Issue, emphasize the importance of checking assumptions that underpin estimation in the context of RCTs and quasi-experimental studies.

---

[3] 'Statistical adequacy' is closely related to the notion of 'congruence', emphasized in the LSE approach (Hendry, 1995, 2009; Bontemps and Mizon, 2003). The common motivation is for the model to be capable of generating data that mimics the observed sample data. However, it is 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: 41), which is consistent with a clearer distinction between statistical and structural/theory models.

[4] Some of the reasons for this are discussed by Spanos (2018, Section 4), who also provides a robust and detailed critique of claims that discourage misspecification testing. Pragmatically, as Gelman (2011: 69) points out, "[m]odel checking plays an uncomfortable role in statistics. A researcher is typically not so eager to perform stress testing, to try to poke holes in a model or estimation procedure that may well represent a large conceptual and computational investment".

Misspecification testing has received very little attention in replication analyses. Consequently, replications that emphasize testing for statistical adequacy would primarily be categorized under replication types that use 'different analysis' compared to the original studies. In terms of Reed's (2018, Figure 1) taxonomy, for example, replications assessing statistical adequacy would therefore usually be classified as 'robustness analyses', including cases where the same data as in the original study, different data from the same population, or data from a different population are used (Reed's replication types 2, 4, and 6, respectively).[5]

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. 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 a key determinant of long-run economic development. It introduces (the logarithm of) historical European settler 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 former 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 the log of settler mortality (with settler mortality capped at a maximum level of 250 per 1,000 people per annum, as in Acemoglu, Johnson and Robinson (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, 2010)). 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

[5] In Reed's taxonomy, a replication assessing statistical adequacy would not generally be classified under 'reproduction', 'repetition' or 'extension' (Types 1, 3, and 5 respectively) unless the original study contained a comprehensive set of relevant diagnostic checks.

considered, including latitude, continental dummies, and dummies for British and French colonies.

AGR report results 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.[6] 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.[7] This testing reveals widespread evidence of model misspecification, with parameter non-constancy and spatial dependence of the residuals being widespread problems. 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 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

---

[6] 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.

[7] 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.

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.[8] 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 to 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: 39, emphasis in original). In IV estimation, the relevant statistical model specified in terms of the observable variables is 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: 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)

[8] 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.

Homoskedasticity	$\text{Var}(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i}) = \boldsymbol{\Omega}$ 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, $\boldsymbol{\Omega}$ 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: 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.[9]

The assumptions in (1)–(3) can be tested using conventional diagnostic tests for normality (Doornik and Hansen, 2008), functional form (Ramsey's (1969) RESET test) and heteroskedasticity (White, 1980). Given the RFs constitute a multivariate linear regression, system misspecification tests, i.e., multivariate equivalents of these single-equation tests (Doornik and Hendry, 2013: 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 RFs and also of break-point Chow tests at different points in the sample (Hendry and Nielsen, 2007: 195–197). Different orderings of cross-sectional data 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.[10]

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

9 See also Spanos (2015: 183; 2018, Section 4.2.2) on the disadvantages of methods that rely on weaker assumptions for their asymptotic properties.

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

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).

In general, there are several issues to consider in the choice of misspecification tests. There is not a `one-size-fits-all' set of misspecification tests that applies to all empirical papers. The choice of relevant misspecification tests will vary depending on the estimation methods used and the nature of the data. It is necessary to test for a variety of different potential departures from a statistically adequate model, but more tests increase the probability of rejection under the null. Multiple testing of different hypotheses can be taken into account, for example, by selecting a numerically smaller significance level for each test (e.g., 1% instead of 5%) in order to control the overall Type I error probability (Hendry and Nielsen, 2007: 135).

For misspecification testing, however, Type II errors are of greater concern than Type I errors. A balance between avoiding spurious rejection of the null of a valid model, while severely testing for potential violations can be achieved by exploiting the different advantages of parametric and non-parametric tests, which are based on different assumptions (Mayo and Spanos, 2011; Spanos, 2000, 2018).[11] The former involve testing against a specific direction of departure under the alternative and have higher local power, whereas the latter have non-directional alternatives (i.e., stating the null is false) and lower local power. Another strategy for balancing the number of tests with coverage of the various alternative departures from the null of a well-specified statistical model is joint testing, using auxiliary regressions that incorporate terms to allow simultaneously for departures from the various assumptions (Spanos, 2000, 2018; Spanos and Mayo, 2015).

In any testing context, fallacies of acceptance and rejection are a concern (Spanos, 2018). Fallacies of acceptance involve interpreting absence of evidence against the null hypothesis as evidence for the null, which, for example, can occur with tests of low power. In contrast, fallacies of rejection occur when evidence against the null hypothesis is interpreted as evidence for the alternative. This can arise due to high-power tests detecting minor violations. It is also a particular problem with parametric misspecification tests as rejection of a specific null does not provide a clear guide to the type of misspecification. For example, rejection of the specific null hypothesis of homoskedasticity does not imply acceptance of the alternative hypothesis that errors are heteroskedastic. Similarly, rejection of the null of parameter constancy can arise for a number of reasons, including outliers, omitted variables or heteroskedasticity. Graphical analysis, motivated as a type of informal severe testing (Spanos, 2006), can be a useful complement for identifying the potential source of misspecification (Spanos, 2018).[12]

Overall, the aim is to combine formal and informal tests to produce an "open-ended exploration devoted to learning the limitations of a fitted model" (Gelman, 2011: 69).

---

[11] 'Severe testing' implies "in cases where the tests have a very high probability of detecting the departures if they were present, … negative misspecification test results provide strong evidence for the absence of any such departures" (Spanos, 2000: 259–260).

[12] In the 'error statistics' approach to inference (e.g., Mayo and Spanos, 2011), severe testing is the key to avoiding the fallacies of acceptance and rejection; this is implemented via a post-data severity evaluation of *p*-values that takes into account the power of the test.

5 What constitutes a 'successful replication'?

Emphasis on probing statistical adequacy as a key component of any replication exercise has important implications for what constitutes a 'successful replication'.[13] The ability to reproduce the reported findings of the original study (e.g., estimated effect sizes, confidence intervals, or *p*-values) would not represent a sufficient condition to be convinced of the trustworthiness of the results if statistical adequacy has not been established. Indeed, the results of multiple replication studies using similar 'curve fitting' procedures applied to different data sets, and reporting similar empirical evidence that appears to support a theory or prior result, may also not be trustworthy, especially if they all neglect investigation of statistical adequacy. Untrustworthy evidence may turn out to be easy to 'replicate' if the methods researchers use ignore statistical misspecifications that are common across studies, as illustrated by Spanos and Mayo's (2015) analysis of generic tests of the Capital Asset Pricing Model.[14]

If significant evidence of misspecification were to be found in the original study, this would potentially point to a 'disconfirmation', and at least flag the need for additional analysis. In such cases, the next step would be to check the extent to which the violations of the statistical assumptions materially affect the results. Depending on the context, some violations may have only minor implications for the properties of estimators and tests, whereas others may have more serious consequences. A 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 the relevant assumptions (e.g., independence of errors, homoskedasticity, etc.) are violated in the data.

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 and provide insights into the robustness of the study's findings. If inferences from a statistically adequate respecified model do not differ markedly from those of the original study, then, although inferences from the original study may not be strictly valid, the interpretation of these results may carry over and the overall conclusions remain robust to the respecification. Alternatively, the statistically adequate respecified model may provide very different results, casting doubt on the reliability of the results of the original study; as an example of the latter, see Akhtaruzzaman, Hajzler and Owen (2018).[15]

---

[13] This terminology, although common, is not ideal. Replications typically yield results along several different dimensions, which deserve a more nuanced, holistic evaluation. Often, this will not be usefully characterized in terms of a binary 'success'/'failure' outcome.

[14] The observation that multiple studies can derive similar results because they use similar methods, but may still be untrustworthy, also has implications for meta-analyses, which conventionally do not seek to place more weight on results from models that pass misspecification tests.

[15] Akhtaruzzaman et al. (2018) replicate a study by Alfaro, Kalemli-Ozcan and Volosovych (2008) which claims that cross-country variation in institutional quality can fully explain the Lucas Paradox, i.e., the tendency for capital to flow mainly to relatively rich countries, contradicting the neoclassical prediction that it should flow to poorer (capital-scarce) countries. Misspecification testing of alternative functional forms of Alfaro et al.'s empirical models demonstrates that their resolution of the Paradox relies on inference in misspecified models. Respecifying the functional form and dealing with outliers yields models that do not fail the misspecification tests but the main

In the AGR study, 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 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 highly 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.

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 data.

Acknowledgements I would like to thank the Editor of the Special Issue, Bob Reed, and three anonymous referees for their detailed, constructive comments and helpful suggestions. This research was supported by a University of Otago Research Grant.

conclusion is reversed: although a country's institutional quality is a quantitatively and statistically significant determinant of capital inflows, its level of per capita income also remains a significant determinant of capital inflows.

References

Acemoglu, D., Gallego, F. A., and Robinson, J. A. (2014). Institutions, human capital, and development. *Annual Review of Economics*, 6(1): 875–912. https://doi.org/10.1146/annurev-economics-080213-041119

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5): 1369–1401. https://doi.org/10.1257/aer.91.5.1369

Acemoglu, D., Johnson, S., and Robinson, J. A. (2012). The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, 102(6): 3077–3110. https://doi.org/10.1257/aer.102.6.3077

Akhtaruzzaman, M., Hajzler, C., and Owen, P. D. (2018). Does institutional quality resolve the Lucas Paradox? *Applied Economics*, 50(5): 455–474. https://doi.org/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(2): 347–368. https://doi.org/10.1162/rest.90.2.347

Angrist, J. D., and Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2): 3–30. https://doi.org/10.1257/jep.24.2.3

Anselin, L., Bera, A. K., Florax, R., and Yoon M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1): 77–104. https://doi.org/10.1016/0166-0462(95)02111-6

Ashraf, Q., and Galor, O. (2011). Dynamics and stagnation in the Malthusian epoch. *American Economic Review*, 101(5): 2003–2041. https://doi.org/10.1257/aer.101.5.2003

Ashraf, Q., and Galor, O. (2013). The "out of Africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 103(1): 1–46. https://doi.org/10.1257/aer.103.1.1

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604): 452–454. https://doi.org/10.1038/533452a

Begley, C. G., and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391): 531–533. https://doi.org/10.1038/483531a

Bontemps, C., and Mizon, G. E. (2003). Congruence and encompassing. In Stigum, B. P. (Ed.), *Econometrics and the philosophy of economics: Theory-data confrontations in economics*. Princeton, NJ: Princeton University Press.

Brown, A. N., and Wood, B. D. K. (2018). Which tests not witch hunts: A diagnostic approach for conducting replication research. *Economics: The Open-Access, Open-Assessment E-Journal*, 12 (2018-53): 1–26. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-53

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5): 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280): 1433–1436. https://doi.org/10.1126/science.aaf0918

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 31(1): 326–342. https://doi.org/10.1111/joes.12139

Cragg, J. G., and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(2): 222–240. https://doi.org/10.1017/S0266466600007519

Doornik, J. A., and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70(s1): 927–939. https://doi.org/10.1111/j.1468-0084.2008.00537.x

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

Doyen, S., Klein, O., Pichon, C.-L., and Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*. https://doi.org/10.1371/journal.pone.0029081

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*. https://doi.org/10.1371/journal.pone.0158064

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*, 107(5): 46–51. https://doi.org/10.1257/aer.p20171031

Easterly, W., and Levine, R. (2003). Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics*, 50(1): 3–39. https://doi.org/10.1016/S0304-3932(02)00200-3

Easterly, W., and Levine, R. (2016). The European origins of economic development. *Journal of Economic Growth*, 21(3): 225–257. https://doi.org/10.1007/s10887-016-9130-y

The Economist (2013). Trouble at the lab. *The Economist*. Available online: https://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

The Economist (2017). Another example of why replication is important in science: Nothing to smile about. *The Economist*. Available online: https://www.economist.com/news/science-and-technology/21731613-nothing-smile-about-another-example-why-replication-important-science.

Edwards, M. A., and Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, 34(1): 51–61. https://doi.org/10.1089/ees.2016.0223

Galiani, S., Gertler, P., and Romero, M. (2017). Incentives for replication in economics. NBER Working Paper No. 23576. http://www.nber.org/papers/w23576

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2: 67–78. http://www.frankfurt-school-verlag.de/rmm/downloads/Article_Gelman.pdf

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics*, 48(3): 283–307. https://doi.org/10.1111/j.1468-0084.1986.mp48003007.x

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., and Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, 9(3): 271–303. https://doi.org/10.1023/B:JOEG.0000038933.16398.ed

Grimes, D. R., Bauch, C. T., and Ioannidis, J. P. A. (2017). Modeling science trustworthiness under publish or perish pressure. bioRxiv. https://doi.org/10.1101/139063

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(1): 83–116. https://doi.org/10.1162/003355399555954

Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics*, 40(3): 715–733. https://doi.org/10.1111/j.1365-2966.2007.00428.x

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054. https://doi.org/10.2307/1912775

Hendry, D. F. (1995). *Dynamic econometrics*. Oxford: Oxford University Press.

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

Hendry, D. F. (2015). *Introductory macro-econometrics: A new approach*. London: Timberlake Consultants.

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 Mills, T. C. and Patterson, K. (Eds), *Palgrave handbook of econometrics: Volume 1, Econometric theory*. Basingstoke: Palgrave MacMillan.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. Thousand Oaks, CA: Sage Publications.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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(21): 2200–2210. https://doi.org/10.1001/jama.2011.713

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2013). Worldwide governance indicators. Methodology and Analytical Issues. World Bank Policy Research Working Paper No. 5430. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1682130

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

Kleibergen, F., and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1): 97–126. https://doi.org/10.1016/j.jeconom.2005.02.011

Mayo, D. G., and Spanos, A. (2011). Error statistics. In Bandyopadhyay, P. S. and Forster, M. R. (Eds), *Philosophy of statistics, Volume 7* (*Handbook of the philosophy of science*). Amsterdam: Elsevier, North-Holland.

McAleer, M. (1994). Sherlock Holmes and the search for truth: A diagnostic tale. *Journal of Economic Surveys*, 8(4): 317–370. https://doi.org/10.1111/j.1467-6419.1994.tb00106.x

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2): 243–251. http://www.jstor.org/stable/2983777

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. https://doi.org/10.17226/21915

Nevo, A., and Whinston, M. D. (2010). Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives*, 24(2): 69–82. https://doi.org/10.1257/jep.24.2.69

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716

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

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(4): 1627–1682. https://doi.org/10.1162/qjec.2010.125.4.1627

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B*, 31(2): 350–371. http://www.jstor.org/stable/2984219

Reed W. R. (2018a). Replication in labor economics. IZA World of Labor. http://dx.doi.org/10.15185/izawol.413

Reed, W. R. (2018b). A primer on the 'reproducibility crisis' and ways to fix it. *Australian Economic Review*, 51(2): 286–300. https://doi.org/10.1111/1467-8462.12262

Sachs, J. D. (2003). Institutions don't rule: Direct effects of geography on per capita income. NBER Working Paper 9490. https://doi.org/10.3386/w9490

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3): 393–415. http://www.jstor.org/stable/1907619

Smaldino, P. E., and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*. http://dx.doi.org/10.1098/rsos.160384

Spanos, A. (1990). The simultaneous-equations model revisited: Statistical adequacy and identification. *Journal of Econometrics*, 44(1-2): 87–105. https://doi.org/10.1016/0304-4076(90)90074-4

Spanos, A. (2000). Revisiting data mining: 'Hunting' with or without a license. *Journal of Economic Methodology*, 7(2): 231–264. http://dx.doi.org/10.1080/13501780050045119

Spanos, A. (2006). Econometrics in retrospect and prospect. In Mills, T. C., and Patterson, K. (Eds), *Palgrave handbook of econometrics: Volume 1, Econometric theory*. Basingstoke: Palgrave MacMillan.

Spanos, A. (2007). The instrumental variables method revisited: On the nature and choice of optimal instruments. In Phillips, G. D. A., and Tzavalis, E. (Eds), *The refinement of econometric estimation and test procedures: Finite sample and asymptotic analysis*. Cambridge: Cambridge University Press.

Spanos, A. (2010). Theory testing in economics and the error statistical perspective. In Mayo, D. G., and Spanos, A. (Eds), *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science*. Cambridge: Cambridge University Press.

Spanos, A. (2015). Revisiting Haavelmo's structural econometrics: Bridging the gap between theory and data. *Journal of Economic Methodology*, 22(2): 171–196. https://doi.org/10.1080/1350178X.2015.1035946

Spanos, A. (2018). Mis-specification testing in retrospect. *Journal of Economic Surveys*, 32(2): 541–577. https://doi.org/10.1111/joes.12200

Spanos, A., and Mayo, D. G. (2015). Error statistical modelling and inference: Where methodology meets ontology. *Synthese*, 192(11): 3533–3555. https://doi.org/10.1007/s11229-015-0744-y

Spolaore, E., and Wacziarg, R. (2009). The diffusion of development. *Quarterly Journal of Economics*, 124(2): 469–529. https://doi.org/10.1162/qjec.2009.124.2.469

Spolaore, E., and Wacziarg, R. (2013). How deep are the roots of economic development? *Journal of Economic Literature*, 51(2): 325–369. https://doi.org/10.1257/jel.51.2.325

Stigum, B. P. (2015). *Econometrics in a formal science of economics. Theory and the measurement of economic relations*. Cambridge, MA: MIT Press.

Stock, J. H., and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Andrews, D. W. K., and Stock, J. H. (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(4): 817–838. http://www.jstor.org/stable/1912934

Woodberry, R. D. (2012). The missionary roots of liberal democracy. *American Political Science Review*, 106(2): 244–274. https://doi.org/10.1017/S0003055412000093