Annette N Brown
Drew B Cameron
Benjamin DK Wood

Quality evidence for policymaking

I'll believe it when I see the replication

March 2014

Replication Paper 1

International Initiative for Impact Evaluation

About 3ie

The International Initiative for Impact Evaluation (3ie) was set up in 2008 to meet growing demand for more and better evidence of what development interventions in low- and middle-income countries work and why. By funding rigorous impact evaluations and systematic reviews and by making evidence accessible and useful to policymakers and practitioners, 3ie is helping to improve the lives of people living in poverty.

3ie Replication Paper Series

The 3ie Replication Paper Series is designed to be a publication and dissemination outlet for internal replication studies of development impact evaluations. Internal replication studies are those that reanalyse the data from an original paper in order to validate the results. The series seeks to publish replication studies with findings that reinforce an original paper, as well as those that challenge the results of an original paper. To be eligible for submission, a replication study needs to be of a paper in 3ie's online Repository of Impact Evaluations and needs to include a pure replication. 3ie invites formal replies from the original authors. These are published on the 3ie website together with the replication study.

The **3ie Replication Programme** also includes grant-making windows to fund replication studies of papers identified on the candidate studies list. Requests for proposals are issued one to two times a year. The candidate studies list includes published studies that are considered influential, innovative or counterintuitive. The list is periodically updated, based on 3ie staff input and outside suggestions. The aim of the 3ie Replication Programme is to improve the quality of evidence from development impact evaluations for use in policymaking and programme design.

About this paper

Quality evidence for policymaking is the inaugural paper for the new 3ie replication paper series. Its authors are all 3ie staff with responsibilities for the replication programme. Authorship is listed in alphabetical order. All content is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or its board of commissioners. Any errors and omissions are the sole responsibility of the authors. Comments and queries should be directed to the corresponding author at bwood@3ieimpact.org.

Suggested citation: Brown, AN, Cameron, DB, and Wood, BDK, 2014. *Quality evidence for policymaking: I'll believe it when I see the replication. Replication Paper 1, March 2014.* Washington, DC: International Initiative for Impact Evaluation (3ie).

3ie Replication Paper Series executive editor: Annette N Brown
Managing editor: Benjamin DK Wood
Assistant managing editor: Jennifer Ludwig
Production manager: Lorna Fray
Assistant production manager and layout: Rajesh Sharma
Copy editor: James Middleton
Cover design: John F McGill
Printer: Via Interactive

Quality evidence for policymaking: I'll believe it when I see the replication

Annette N Brown
Drew B Cameron
Benjamin DK Wood

3ie Replication Paper 1
March 2014

International Initiative for Impact Evaluation

Acknowledgements

The International Initiative for Impact Evaluation funded this research. Sebastian Insfrán Moreno provided valuable research assistance. We are grateful to Beryl Leach, Bruce McCullough, Richard Palmer-Jones and Heather Lanthorn for helpful comments. We thank Michell Dong for formatting assistance.

Abstract

In this paper, we make the case for replication as a crucial methodology for validating research used for evidence-based policymaking, especially in low- and middle-income countries. We focus on internal replication or the re-analysis of original data to address an original evaluation or research question. We review the current state of replication in the social sciences and present data on the trends among academic publications. We then discuss four challenges facing empirical research that internal replication can help to address.

We offer a new typology of replication approaches for addressing these challenges. The types – pure replication, measurement and estimation analysis, and theory of change analysis – highlight that internal replication can test for consistency and statistical robustness but can and should also be used to ensure that a study fully explores possible theories of change in order to draw appropriate conclusions and recommendations for policymaking and programme design.

JEL C18 D04 O20

Keywords: replication, statistical methods, theory of change, policy analysis, impact evaluation

Contents

Acknowledgements i
Abstract ii
List of figures and tables iii
Abbreviations and acronyms iv
1. Introduction 1
2. Replication policies and practice in social science 2
3. Why replication is necessary 6
3.1 Challenge 1: to err is human 6
3.2 Challenge 2: it is not a perfect science 8
3.3 Challenge 3: publish or perish 9
3.4 Challenge 4: policy recommendations please 10
4. Addressing the challenges: approaches to replication 11
4.1 Pure replication 12
4.2 Measurement and estimation analysis 13
4.3 Theory of change analysis 15
5. Other types of replication 16
6. Conclusion 17
References 19
Appendix A: Journal replication policy survey results 25

List of figures and tables

Figure 1: Top economics and development publications' replication policies 4

Abbreviations and acronyms

3ie	International Initiative for Impact Evaluation
AIDS	Acquired immune deficiency syndrome
FDIC	Federal Deposit Insurance Corporation
HIV	Human immunodeficiency virus
IPA	Innovations for Poverty Action
J-PAL	Abdul Latif Jameel Poverty Action Lab
JSY	Janani Suraksha Yojana
MEA	Measurement and estimation analysis
NGO	Non-governmental organisation
PEPFAR	President's Emergency Plan for AIDS Relief
PhD	Doctorate of Philosophy
RCT	Randomised controlled trial
TCA	Theory of change analysis

1. Introduction

Every so often, a well-publicised replication study comes along that, for a brief period, catalyses serious discussion about the importance of replication for social science research, particularly in economics. The most recent example is the Herndon, Ash, and Pollin replication study (2013) showing that the famous and highly influential work of Reinhart and Rogoff (2010) on the relationship between debt and growth is flawed.

McCullough and McKitrick (2009) document numerous other examples from the past few decades of replication studies that expose serious weaknesses in policy influential research across several fields. The disturbing inability of Dewald *et al.* (1986) to replicate many of the articles in their *Journal of Money, Credit and Banking* experiment is probably the most well-known example of the need for more replication research in economics. Yet, replication studies are rarely published and remain the domain of graduate student exercises and the occasional controversy.

This paper takes up the case for replication research, specifically internal replication, or the reanalysis of original data to address the original evaluation question. This focus helps to demonstrate that replication is a crucial element in the production of evidence for evidence-based policymaking, especially in low- and middle-income countries.

Following an overview of the main challenges facing this type of research, the paper then presents a typology of replication approaches for addressing the challenges. The approaches include pure replication, measurement and estimation analysis (MEA), and theory of change analysis (TCA). Although the challenges presented are not new, the discussion here is meant to highlight that the call for replication is not about catching bad or irresponsible researchers. It is about addressing very real challenges in the research and publication processes and thus about producing better evidence to inform development policymaking.

Replication for development impact evaluations

Replication to validate policy-relevant findings is important for all research that is used to inform policy and practice. In the case of impact evaluations for development, internal replication is even more important: first, because single studies can strongly influence policy; and second, because external replications – where the intervention is conducted again in the same or similar contexts – are difficult and extremely rare.

When single evaluations are influential, and any contradictory evaluations of similar interventions can be easily discounted for contextual reasons, the minimum requirement for validating policy recommendations should be recalculating and re-estimating the measurements and findings using the original raw data to confirm the published results, or a pure replication. A more comprehensive internal replication, for example, one that includes robustness checks on the published findings, goes even further to validate policy recommendations.

One example of a single study that has been highly influential is Miguel and Kremer (2004) on the benefits of deworming programmes for children, which is the primary motivation for the Deworm the World initiative that promotes deworming in schools around the world.

Cattaneo *et al.* (2009), in an evaluation of an early round of the *Piso Firme* project in Coahuila state for fewer than 3,000 households, find that cement flooring has dramatic impacts on child health and cognitive development. The results helped influence the Mexican government to scale up the project nationally and further fuelled international efforts promoting cement floors through the non-governmental organisation (NGO) Un Techo para mi País, which operates in 19 Latin American countries.[1]

And just three evaluations of medical male circumcision (Bailey *et al.* 2007; Gray *et al.* 2007; Auvert *et al.* 2005) in Africa made male circumcision a core strategy for HIV prevention for prominent funders, including the President's Emergency Plan for AIDS Relief (PEPFAR) and the Bill & Melinda Gates Foundation.

Some argue that the referee process is responsible for ensuring the quality of the results that are ultimately published, and thus replication is not necessary. The famous Sokal experiment (Sokal and Bricmont 1998) demonstrates that the peer review process does not ensure quality in humanities journals; and more recently, Bohannon demonstrates that the peer review process does not ensure quality in science journals (*The Economist*, 5 October 2013, p.85).

In both experiments, the scientists submitted noticeably bogus articles to scholarly journals and found that the referees and editors accepted them for publication in most cases. Few, if any, referees of empirical articles request the data and re-estimate the models in the papers, which leaves editors to rely primarily on authors' credentials and the general feasibility of results in order to determine the validity of the recommendations. Hamermesh (2007 p.720) identifies this problem for economics. *The Economist* (19 October 2013 p.28) more recently points out the same problem for the sciences.

2. Replication policies and practice in social science

Repeated calls for increased replication in social science usually fall on deaf ears.[2] A slew of publications advocate for standardising internal validation through replication (Collins 1984; King 1995; Falk 1998; Abbott 2007; Evanschitzky *et al.* 2007; Freese 2007; Valentine *et al.* 2011) and specifically in economics (Mittelstaedt and Zorn 1984; Dewald *et al.* 1986; Hubbard and Vetter 1992; Hamermesh 2007; Burman *et al.* 2010). Folbre (2013) is the latest in a long line of economists to highlight the need for more replication research, in the wake of Herndon and others' (2013) replication findings of Reinhardt and Rogoff. But such prescriptions have led to only limited institutional change in the discipline.

[1] Funding efforts for the international *Un Techo para mi País* programme have expanded based on *Piso Firme* evaluation: http://cega.berkeley.edu/assets/cega_events/19/E2A_Cement_Floors_Brief.pdf.
[2] More general requests for replication in the social sciences have been made, for example, by King (1995) and Yong (2012).

Around the turn of the century, three journals – *Journal of Political Economy*, *Empirical Economics* and *Labour Economics* – attempted to promote and publish replications but their efforts were generally short lived because of a lack of interest (Hamermesh 2007 p.723). Similarly in 2003, the editors of several international relations journals[3] signed a minimum standards requirement, which states that accepted authors of empirical articles must make their data publicly available for replication (Bueno de Mesquita *et al.* 2003 p.105). Although the number of data replication policies appears to be on the rise, Gleditsch and Metelits (2003 p.76) conclude that 'journals [with] a replication policy often fail to implement or enforce it'.

In 2004, the American Economic Association established a replication policy for its premier journal the *American Economic Review.* It states, 'It is the policy of the *American Economic Review* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication' (Bernanke 2004 p. 404).

The policy has since been strengthened with systematic enforcement, alternative means of access to data for those requesting exemption and the establishment of a team of graduate students that conducts checks of all submitted files for compliance with the policy. McCullough (2009 p.122) reports that several of the top economics journals, for example, the *Journal of Political Economy* and *Econometrica*, have adopted mandatory data and code policies, but notes that the profession is 'a long, long way from ensuring that most of its published results are replicable.'

To see whether the prevalence of journal replication policies has changed in the past seven years, we conducted a virtual and actual survey of the top 50 economics journals along with 15 additional international development journals.[4] The virtual survey gathered information about policies published on the journals' websites. Where there was no information, we requested it through email and phone correspondence.

The surveyed journals are listed in Appendix 1. The journals' replication policies are grouped into five categories:

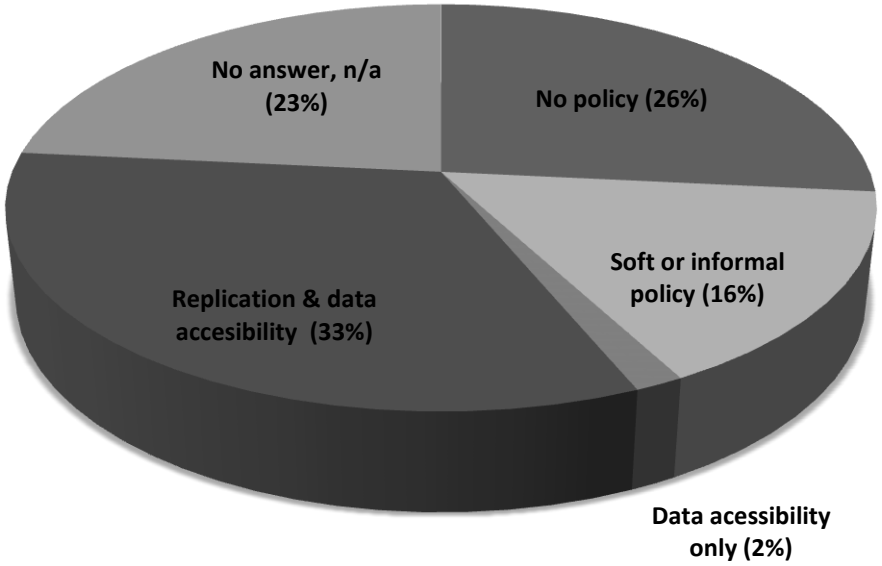
- confirmed to have no replication policy
- not applicable (does not publish original research) or no answer to repeated inquiries about replication policy
- promotes replication as an important practice but has merely a soft or informal policy

[3] *The Journal of Peace Research*, *International Studies Quarterly*, *International Interactions*, and the *Journal of Conflict Resolution*.

[4] Our economics journal rankings came from IDEAS/RePEc (http://ideas.repec.org/top/top.journals.simple.html) on 12 July 2013, using their simple ranking method. Our selection of the 15 development journals was informal but guided by an editor for the *Journal of Development Studies* and the *Journal of Development Effectiveness.* The *Journal of Development Economics* overlapped both categories making the final sample 64 journals. For the complete list of surveyed journals please see Appendix 1.

- has a data accessibility policy with no mention of replication-ready data
- has a robust replication policy and standards for data accessibility (see Figure 1).

Figure 1 Top economics and development publications' replication policies



Note: 64 publications surveyed

According to our survey, roughly half of the journals recognise the importance of replication. Twenty-one (33 per cent) of the 64 peer-reviewed journals have an active, formal replication policy. Only one journal has a data accessibility-only policy. Ten journals (16 per cent) have a soft or informal replication policy that encourages the practice. The other 32 journals (49 per cent) have no specific replication policy, are not applicable or did not answer repeated requests for clarification.[5] The survey reveals that the American Economic Association is a leader in promoting replication policies.[6]

These developments are promising but requiring replication data to be made available is only one side – the supply side – of increasing the practice of internal replication. The constraints on the demand side – for replication researchers, original authors and journal editors – continue to be the same as Hamermesh (2007), Abbott (2007) and others describe.

[5] Correspondence with the *World Bank Economic Review* editors led to a very encouraging response. Initially, they responded to our inquiries by indicating that they had no specific replication policy. Soon after, we received an email informing us that, after an internal review, they had decided to adopt a new replication policy.
[6] The *American Economic Review* is cited as a model by multiple other publications for its exemplary replication policy, including the *Journal of Political Economy*, the *Journal of Labor Economics*, the *Journal of Economic Perspectives*, and *Brookings Papers on Economic Activity, Economic Studies Programs*.

For replication researchers, a replication study that confirms the findings of the original article is rarely given much attention. When replication studies are published, they often appear in the 'comments' section of the original journal or in a less prominent journal than the one that published the original article.[7] Journal editors also have a disincentive to publish replications that refute articles published in their own journals. Indeed, high-profile refutations might diminish an entire journal's standing (Abbott 2007 p.215).

Original authors usually have the most to lose and the least to gain from a replication study of their work, causing them to resist requests for data and code from replication researchers. In order to facilitate a successful replication, original authors often need to invest time in documenting work they completed years ago and/or compiling data into files that can be easily transferred and understood.

A successful replication – one that validates the original results – should build an original author's reputation, but that only works to the extent that successful replications are published and publicised. A replication that refutes the findings or policy recommendations of the original study may not just call into question the original study but also other studies by the same original authors.

Furthermore, the process of obtaining data often suffers delays because many original authors want to develop as much original scholarship as possible from a data-set before making it public (Abbott 2007 p.212). Even when data are provided, reconstructing the estimation data and reproducing the original results can be enormously time consuming, particularly if the original article was shortened for publication, leaving out important information about how variables were created or data were transformed. As *The Economist* (19 October 2013 p.28) points out, 'replication is hard and thankless'.

On the bright side, there is some evidence of a correlation between public data availability and increased citation counts in the social sciences. Gleditsch (2003) finds that articles published in the *Journal of Conflict Resolution* that offer data in any form receive twice as many citations as comparable papers without available data (Gleditsch *et al.* 2003; Evanschitzky *et al.* 2007).

And although only 18 out of 120 political science journals that Gherghina and Katsanidou (2013 p.12) identify have a public replication policy (that accepted authors are – or may be – required to submit data and associated files), 'journals with more citations [are] more likely to have a data availability policy than publications with fewer citations.' Although the direction of causality is unclear (whether increased citations are a result of more articles being published that use that data, or whether the very availability of data leads to greater recognition of article in question), original authors should find comfort in the notion that public data are likely to improve an article's impact.

[7] For example, almost all of the replication studies cited by Hamermesh are published as comments. A further case in point is Hamermesh's 2007 article on replication, which appeared in *Canadian Journal of Economics*, though Hamermesh's other research is routinely published by journals such as *American Economic Review* and *Review of Economics and Statistics*.

Despite the well-known constraints, the practice of replication is on the rise. A primary source of replications today is the widespread use of internal replication as an exercise to learn methods in graduate student courses. The Herndon *et al.* (2013) replication originates from a class assignment, for example.

These replication studies, though, often focus primarily, if not exclusively, on pure replication. Students replicate the results of the original paper in order to learn how the original authors applied the methods to the data, not necessarily to validate the robustness of the policy conclusions. An exception is the advice that King (2006) gives to students to encourage them to get their replication studies published, but the full extent of his advice actually exacerbates reporting and publication bias, as discussed further below.

With funding from the Institute for New Economic Thinking, the Centre for Statistics at the University of Göttingen, Germany has begun its own replication working paper series, encouraging graduate students in PhD seminars, as well as others, to conduct pure replications of empirical work. The initial papers of these replication studies (Wohlfarth 2012; Zakula 2012) have already been published online in the working paper series and future seminars and a wiki site for idea exchange are planned. The International Initiative for Impact Evaluation's (3ie) Replication Programme similarly encourages replication of development impact evaluations that have strongly influenced policy by funding replication studies through a small grants window, providing limited data preparation support to original authors, and publishing a paper series.

3. Why replication is necessary

Despite some evidence of codified replication policies and a growing conceptual interest in the practice, replication remains an under-used tool in social science and international development. The primary thesis of this paper is that internal replication is necessary to establish the credibility of empirical findings used for policy decisions due to very real constraints in the research and publication processes.

Replication should be seen as part of the process for translating research findings into evidence for policy and not as a way to catch or call out researchers who, in all likelihood, have the best of intentions when conducting and submitting their research, but face understandable challenges. These challenges include the inevitability of human error, the uncontrolled nature of social science, reporting and publication bias, and the pressure to derive policy recommendations from empirical findings. We discuss these four challenges below.

3.1 Challenge 1: to err is human

People make mistakes. For this reason alone replication is a valuable tool. Economics has dealt with its fair share of human errors over the years, which should further justify the need for increased replication research. Feldstein (1974) on how the US social security programme affects personal savings is a good example of how an inadvertent error may alter research findings. Leimer and Lesnoy (1982), in their replication study, discover a programming error that both

reduces the magnitude of the coefficient and enlarges its standard error. Feldstein (1982), a renowned Harvard University professor and president *emeritus* of the National Bureau of Economic Research, graciously acknowledges the error and discusses its significance.

More recently, Donohue and Levitt (2001) argue that legalising abortion in the United States reduces crime rates. Foote and Goetz (2008) uncover a coding error in the original research, demonstrating that the original authors do not actually control for interaction effects as they claim. Donohue and Levitt (2008) acknowledge the mistake when re-estimating the results in response to the replication study.

In the development sphere, Iversen and Palmer-Jones in their forthcoming replication study of Jensen and Oster (2009), uncover a coding error and notified the original authors prior to the publication of the replication study. Jensen and Oster subsequently acknowledge this error in their corrigendum (2012). There are many more examples of innocent errors, though not so many examples of gracious acknowledgement.

The line between innocent error and known distortion or deception can be hard to draw. In a *New York Times* Economix blog on the oopsies in economics studies, Rampell (2013) reports that errors are often blamed on 'the poor research assistant who did the grunt work.' It is easy to imagine that an innocent error could turn into a known distortion when a researcher, on realising the error, seeks to cover it up or defend it on the grounds of time and reputational considerations. In any discipline, misconduct is sure to occur when those at the margins take calculated risks in favour of material well-being and professional advancement over the precision of scholarship (Wible 1992 p.20).

The Economist explores the problem of errors in the sciences. It cites professional pressure, competition and ambition as barriers to the self-correction mechanism that should exist in the scientific process to reduce or correct errors over time. It concludes, 'There are errors in a lot more of the scientific papers being published, written about and acted on than anyone would normally suppose, or like to think' (*The Economist* 19 October 2013 p.26).

The role played by replication then is to both find and correct innocent errors and to change the calculations of researchers who do not check their own (or their research assistants') work, or who do find errors but would face significant costs to correct them. Without replication, the cost of correcting an error may be the possibility that the research, now with insignificant or different findings, would not be published. With replication, although a researcher may emerge untarnished the first time he or she graciously acknowledges an error uncovered by replication, after the second or third time replication studies uncover errors, there is likely to be a reputational consequence. With a high enough probability of replication, incentives, and thus practices, should change.

3.2 Challenge 2: it is not a perfect science

Even accounting for the recent popularity of randomised controlled trials (RCTs) in the social sciences, social science empirical research is not like the medical and natural sciences. In a medical efficacy trial, the focus is on precisely determining and controlling the conditions of the trial so that the result is as simple as a comparison of the observed outcomes. For such trials, validation comes from external replication – a new trial is conducted on a newly drawn sample of patients – rather than from recalculating the comparison of the outcomes.

In most social science empirical research, much of the focus is on statistical methods and the assumptions needed to justify the use of certain methodologies. The assumptions researchers make, the indicators they select or create to measure social and economic concepts, and the estimation methods they employ, are all human choices and not controlled lab conditions or biological and physical properties.

One only needs to read a sampling of social science articles over time on a particular question or programme to see that econometrics and statistics are not perfect sciences. The debates about methodology are rarely neatly resolved to everyone's agreement. Leamer (1983) discusses this problem in his famous article, 'Let's take the con out of econometrics'. He provides a compelling example of the various models that can be estimated to establish the effect of capital punishment on murder rates, from which he concludes that 'any inference from these data…is too fragile to be believed' (Leamer 1983 p.42). He concludes that 'in order to draw inferences from data as described by econometric texts, it is necessary to make whimsical assumptions' (*ibid*. p.43).

In response to Angrist and Pischke's article, 'The credibility revolution in empirical economics', Leamer (2010) updates his 1983 article and maintains his case that econometric inference is often fragile. Certainly, advances have been made in techniques but there is still very little sensitivity analysis. Published articles only present the preferred specifications and the sensitivity analyses that the authors choose to present. Leamer (2010 p.32) beseeches, 'Can we economists agree that it is extremely hard work to squeeze truths from our data-sets…?'

Advocates of RCTs argue that social science trials come closer to real science where the findings are not subject to the vagaries of statistics. Rarely, however, do RCT-based studies report only the comparison of the treatment and control means for just the groups subject to random assignment – the basis for the claim that there is no selection bias. As Deaton (2010 p.447) argues:

> …conducting good RCTs is exacting and often expensive, so that problems often arise that need to be dealt with by various econometric or statistical fixes. There is nothing wrong with such fixes in principle…but their application takes us out of the world of ideal RCTs and back into the world of everyday econometrics and statistics. So that RCTs, although frequently used, carry no special exemption from the routine statistical and substantive scrutiny that should be routinely applied to any empirical investigation.

Even in the medical sciences, the analysis of heterogeneity of outcomes, or post-trial subgroup analysis, is not accorded 'any special epistemic status' by the United States Food and Drug Administration rules (Deaton 2010 p.440). In the social sciences, testing for and understanding heterogeneous outcomes is crucial to policymaking. An average treatment effect demonstrated by an RCT could result from a few strongly positive outcomes and many negative outcomes, rather than from many positive outcomes, a distinction that would be important for programme design.

Most RCT-based studies in development do report heterogeneous outcomes. Indeed, researchers are often required to do so by funders who want studies to have policy recommendations. As such, RCTs as practised – estimating treatment effects for groups not subject to random assignment – face the same challenges as other empirical social science studies.

3.3 Challenge 3: publish or perish

It is no great secret that researchers across many disciplines are incentivised to report statistically significant results in their published work, whether from their own desire to make their work more compelling or in response to editors' desires to publish interesting articles. Publication bias – here encompassing both researcher reporting bias and editor publication bias – is well documented in the literature (Lipsey and Wilson 1993; Ioannidis 2005; Fanelli 2010; Yong 2012).

Publication bias means that published research may be systematically unrepresentative of populations under study (Rothstein *et al*. 2005). Furthermore, such bias causes what many have called the 'file drawer problem', the extreme of which would be that 'journals are filled with the 5% of studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant results' (Rosenthal 1979 p.638).

Gerber and Malhotra (2008) examine the evidence for publication bias. They review 13 years of statistical articles published in *American Political Science Review* and *American Journal of Political Science* and look at the ratio of results right above and right below a 0.05 *p*-value. In the absence of publication bias, the ratio should be one; the distribution around the arbitrarily chosen 0.05 *p*-value should be smooth. Instead, Gerber and Malhotra find dramatically more reported results above the critical value than below it. Their histograms show spikes of reported values at the critical points.

Humphreys *et al.* (2013) provide a good overview of the 'fishing' problem, whereby researchers fish for interesting results in their data. Fishing is an extreme version of reporting bias: not only do researchers choose to report only interesting results, they also adapt their models and specifications in order to yield statistically significant results.

The solution most often recommended for the publication bias challenge is research registration. Prospective registration of hypotheses or, even better, of entire analysis plans as Humphreys *et al.* (2013) argue, helps to address reporting bias by giving referees and readers a way to check that results are

reported against intended tests. Research registries also help to address publication bias by giving systematic reviewers and others a way to see what studies might have been started, or even finished, but never published.

Registries are only part of the solution though. Most do not require submission of a complete analysis plan, so registrants still have quite a bit of latitude in what they report beyond the basic hypotheses entered into the registration form. Registration is also quite new in the social sciences. Even as journals and funders start to require registration, it will be years before the majority of published articles will have a public registration on file.

Replication is another way to test an article for reporting and publication bias. Replication cannot uncover studies that were never published, but it can explore how selective the reported results seem to be, whether due to the author's selection or the referee's or editor's selection. Ironically, publication bias extends to the publication of replications. As discussed above, journal editors typically want to publish articles that are new and so relegate replication studies to comments sections at best.

King (2006) encourages graduate students to conduct replication studies but, in his desire to help students publish, he suggests they may leave out replication findings that support the original article and instead look for findings that contribute by changing people's minds about something. About sensitivity analysis, King (2006 p.121) advises, 'If it turns out that all those other changes don't change any substantive conclusions, then leave them out or report them very briefly.' While this advice is understandable for helping students to publish, it limits the role that replication can and should play in identifying those published results that are highly credible and therefore most useful for policymaking, in addition to identifying those results that are perhaps subject to fishing.

3.4 Challenge 4: policy recommendations please

The concerted push for the statement of policy recommendations, particularly from research in international development, can create perverse incentives for researchers in the analysis and reporting of their research. Research sponsors such as 3ie, the Abdul Latif Jameel Poverty Action Lab (J-PAL), and Innovations for Poverty Action (IPA) and Evidence Action have explicit objectives to translate research into policy. J-PAL's mission is 'to reduce poverty by ensuring that policy is based on scientific evidence, and research is translated into action.'[8] 3ie publicly states its preference for greater policy influence and policy relevance in its selection criteria for impact evaluation awards.[9]

Journals also emphasise the importance of policy recommendations, particularly those journals designed to publish applied research. A review of the submission criteria for the websites of the top 15 journals in international development[10] reveals varied emphasis on providing policy recommendations for submitting authors. More than half of development journals mention the promotion of policy

[8] See J-PAL's mission statement at: http://www.povertyactionlab.org/about-j-pal.
[9] 3ie rewards research proposals deemed to have greater policy impact: http://www.3ieimpact.org/en/funding/open-window/ow-faq/#35.
[10] Derived from the list of development journals in Appendix 1.

relevance. *The Journal of Development Effectiveness* emphasises the 'use of findings to inform policy and program design' and *Development Policy Review* seeks to 'extend or challenge the leading policy themes of the day … [and] speak to practical policy problems and frameworks.'

The emphasis on policy recommendations is laudable in the quest to improve evidence-based policymaking. *Ex ante,* policy relevance considerations should lead to better designed studies, which is why research sponsors emphasise policy relevance in their funding competitions. *Ex post,* however, particularly in the absence of *ex ante* publication of comprehensive analysis plans, the push for policy recommendations may lead researchers to draw policy conclusions consistent with, but not proven by, their study's findings. Even when researchers are careful not to overstate their policy conclusions – think about how many papers conclude with 'more research is needed' – others can be quick to make policy recommendations based on the tested (or implied) theory of change without asking whether alternative theories, or different causal mechanisms, were also tested.

Replication can provide the opportunity to further explore the causal chain using the article's own data and perhaps adding data and information from other sources. A replication study can be used to conduct sensitivity analysis on the policy recommendations in much the same way as it can be used to conduct sensitivity analysis on the primary estimates.

4. Addressing the challenges: approaches to replication

Social scientists have created numerous typologies around replication. In this section, we focus primarily on typologies from the economics literature.[11] In early debates, economists expressed epistemological concerns over the notion of replication that have led to a call for distinction between routine checking of results (verifying the instruments), replication (doing the same experiment again with the same instruments) and reproduction (conducting a new experiment) of scientific findings (Cartwright 1991).

Hamermesh (2007) proposes a different typology, which focuses on the underlying data used to test a model. In Hamermesh's parlance, pure replication is the duplication of the statistical experiment. Statistical replication is conducting the same statistical experiment on a different sample drawn from the same population, which he rightly suggests would be rare in economics because economists and other empirical social sciences typically use all the data available to them in the original study. Scientific replication is when the statistical experiment is conducted on a different sample drawn from a different population and perhaps with an altered model. Hamermesh suggests that most of the replication in economics falls under this third definition.

[11] For more information on replication research in: political science see King (1995) and Herrnson (1995); marketing see Berthon *et al.* (2002), Evanschitzky *et al.* (2007), Toncar and Munch (2010); public health see Valentine *et al.* (2011); and experimental psychology see Tweney (2004) and Binmore and Shaked (2010).

García (2013) develops a Bayesian framework for replication studies. His framework defines five types of replication study: pure, external, robustness, statistical and procedural. He defines pure and statistical replication the same way as Hamermesh. External replication is repeating the experiment on a new population. Robustness replication checks the sensitivity of the findings against the decisions made by the researchers in their analysis – what García calls the 'researchers' opinions'. Procedural replication, which García advocates, replaces 'the inferred standard from a pure replication with generally accepted standards whenever the two differ' (2013 pp.11–12).

We propose a typology for approaches to internal replication – replication using the data from the original study – that focuses attention on the types of robustness checks and model alterations that would fall under Hamermesh's scientific replication and García's robustness replication. Three distinctions can be made that roughly match the replication approaches to the challenges in research and publication that authors face.

First is pure replication, which is using an author's original data and methodologies to reproduce the published results. Second is MEA, which is using alternative measurement and estimation techniques to examine the same questions posed by the original authors. Third is TCA, which is the examination of alternative theories of change using the same data. MEA and TCA may also incorporate new data. We discuss these three approaches in turn.

4.1 Pure replication

Pure replication is the reproduction of the original study results using the original data – as Hamermesh (2007 p.716) writes, '…examining the same question and model using the underlying original data-set….' This exercise is important for validating the original results and is also the necessary first step for further replication tasks. Pure replication includes reconstructing the estimation variables, rewriting and rerunning programmes for the estimations and auditing the original data manipulation and estimation code, particularly when the replication results differ substantively from the originals.

Pure replication appears straightforward on the surface, but published articles rarely include all the information needed to replicate the tables from the original data. Working paper versions of published articles can help, as they are often more comprehensive. But the process can still be very time consuming and may necessitate communicating with the original authors. King (2006 p.120) provides a useful step-by-step guide for students conducting pure replications but notes, 'Replicating an article, even if you secure access to the original data, is normally a highly uncertain and difficult process.'

Pure replication addresses Challenge 1: to err is human. There is no reason to believe that any particular published study is free from unintentional error. Pure replication therefore plays a vital role in validating the results of a study that policymakers will use. Of course, replication researchers may make mistakes as well, but experience shows that original authors are usually eager to conduct the replications to uncover them, so that ultimately the process reduces the total number of errors.

4.2 Measurement and estimation analysis

MEA builds on pure replication to further test the robustness, or in Leamer's words, 'sensitivity' of the original findings beyond the checks that the original article employed. MEA incorporates aspects of Challenge 2: it is not a perfect science and Challenge 3: publish or perish. Original authors may not have tested plausible alternative specifications and/or reporting bias may have resulted in alternative specifications going undocumented.

As the label suggests, MEA can involve analysis of the sensitivity of measurement, analysis of the sensitivity of estimation or both. Examples of MEA include redefining and recalculating the variables of interest, introducing additional control or interaction variables, and using alternative estimation methodologies. MEA should not be a data mining exercise; the robustness checks for the replication study should be planned and justified in advance.

There are many examples in economics where variable measurement makes a difference. Researchers concerned with individual and household welfare have grappled for decades with whether income or consumption measures provide the most accurate reflection of welfare and poverty. More recently, the focus has turned to measuring wealth, typically through asset-based strategies.

Carter and Barrett (2006) explain the importance of accounting for assets when measuring persistent poverty.[12] Basu and Foster (1998 p.1746) propose an alternative measurement for household literacy, demonstrating how this new measure changes the picture of literacy in India and argue that 'changing the way literacy is measured is likely to alter the perceived efficacy of actual literacy programmes, which in turn may influence their design.'

Measurement analysis (the 'M' in MEA) is important any time an original study employs an author-constructed index, especially an index created from separate qualitative variables. Iversen and Palmer-Jones (forthcoming) in their replication study of Jensen and Oster (2009) examine two key index variables from Jensen and Oster – female autonomy and attitudes towards spousal violence.

Jensen and Oster construct each index variable by combining six different qualitative variables captured in the survey. Iversen and Palmer-Jones examine the individual variables separately, test different measurements for those variables, and also explore the robustness of the variables against those constructed from similar questions in another data-set.

They find that the different elements of the indexes respond differently to the intervention (introduction of cable television). Examining these differences, particularly in light of theories of female empowerment, enriches the understanding of how cable television could affect attitudes, though Iversen and Palmer-Jones's measurement analysis of these variables does not materially change the Jensen and Oster findings.

[12] The literature has expanded alternative wealth measurements to include expenditure and consumption, including Chen and Ravallion (2007) and Liverpool-Tasie and Winter-Nelson (2011).

The famous Feldstein (1974) article cited above provides another example of the difference measurement can make. In a 1995 update of his original study of social security and savings, Feldstein concludes that his original results hold when tested with additional years of data. Baker and Rosnick (2013) discuss their attempt at a pure replication of the new study and their inability to replicate Feldstein's results. Correspondence between Baker and Feldstein reveals that the results are not replicable because the social security administration recalculated one of the key variables in the data.[13]

Estimation analysis is quite common in economics, in the sense that many researchers challenge the findings of earlier research by arguing that the estimation was wrong, or could be improved, and then conducting new estimations with the same data. When published though, these re-estimation studies typically do not include a pure replication and are not identified as replication studies.

The debate over the effectiveness of India's Janani Suraksha Yojana (JSY) conditional cash transfer programme for increased health facility use during pregnancies provides a good example of replication analysis that is not labelled as such. The original authors, Lim *et al.* (2010), evaluate the programme and find decreasing neonatal mortality rates as a causal result of the JSY programme. However, they also determine that the programme's pro-poor targeting is ineffective, with a significant amount of leakage to middle-class mothers.

Mazumdar *et al.* (2011) argue that Lim and others' (2010) approach is invalid because their matching estimation strategy is unable to control for unobserved heterogeneity and question the causal interpretations because of the possibility of reverse causality. Mazumdar *et al.* use a difference-in-differences alternative estimation strategy on the same data used in the original Lim *et al.* study to show that JSY targets the poor well but does not reduce neonatal mortality.[14]

Going back to the US economy, a Boston Federal Reserve Bank (Boston Fed) study provides excellent examples of replication studies that employ MEA and find dramatically different results from the original study by Munnell *et al.* (1992). The bank released a working paper in 1992 that shows that race was a significant factor in lending and led to widespread rule changes in lending practices (McCullough and McKitrick 2009). Four years later, Munnell *et al.* (1996) published a follow-up version of the same paper in the *American Economic Review*.

McCullough and McKitrick (2009) provide a detailed description of two replication studies, one that focuses on measurement analysis and the other on estimation analysis. Day and Liebowitz (1998) start with information uncovered by an employee of the Federal Deposit Insurance Corporation (FDIC) that in 26 cases the Boston Fed authors classify applicants as rejected when that was not the case

[13] For more information on Baker's replication attempt see: http://www.cepr.net/index.php/blogs/beat-the-press/in-history-of-economic-errors-martin-feldstein-deserves-mention.
[14] Rokicki and Carvalho's (forthcoming) 3ie-funded replication study will shed more light on robustness of the original impact evaluation.

(Horne 1994).[15] They also note that the Boston Fed authors use alternative indicators to measure creditworthiness instead of the credit score that the bank calculates. When Day and Liebowitz change the 26 cases back and use the bank's credit score, they find that the estimated effect of discrimination goes to zero.

Harrison (1998) revisits the full Boston Fed data-set and finds that there are additional variables available that are not included in the Boston Fed estimations. Examples are marital status, age, and the verity of the application information. Harrison re-estimates the equations using these variables and finds that they are statistically significant and that the estimated effect of discrimination goes to zero when they are included.

4.3 Theory of change analysis

In TCA, replication researchers explore the original research from a TCA perspective. It is intended to provide the users of evidence with a better understanding of the causal pathway, or pathways, underlying the studied intervention. TCA relates to Challenge 4: policy recommendation please, because policy analysts may infer significant policy implications from a study, even if the study does not provide evidence at all points of the causal chain or explore the possibility of equally valid alternative hypotheses. As with MEA, TCA should be planned and justified in advance.

Iversen and Palmer-Jones (forthcoming) re-examine Jensen and Oster's (2009) influential study on 'The Power of TV: Cable Television and Women's Status in Rural India'. Although Iversen and Palmer-Jones generally confirm the original research results, they check for robustness to alternative variable constructions and extend the analysis by testing for heterogeneous outcomes in order to uncover what causal mechanisms might be at play. They find, for example, that cable television access only increases female autonomy for educated women, which suggests that the underlying mechanism is more complicated than the original study implies. This replication result may attenuate the social benefits expected by the Indian state of Tamil Nadu, which had planned to provide free colour televisions to all households.

Another example of a replication study that focuses on the theory of change is the Cervellati *et al.* (2014) replication study of the highly influential work of Acemoglu *et al.* (2008). Both studies explore Lipset's (1959) 'modernization hypothesis' that increasing per capita income in developing countries leads to increased democracy.

Acemoglu *et al.* show that the previously estimated positive relationship disappears when country and fixed effects are included. Using the original authors' data, Cervellati *et al.* first conduct a pure replication and MEA, and find that the original study's results are robust. They then explore the underlying theory more carefully and posit that colonisation affects both the income paths and democracy adoption in the studied countries. They examine this hypothesis by testing for heterogeneous effects by colonisation and find that colonisation

[15] See also Horne (1997).

indeed makes a difference. Non-colonised countries display a positive relationship between income and democracy but colonised countries display a negative relationship.

Both MEA and TCA can employ data from other sources. In MEA, for example, the replication researcher may test the model on standardised test scores from administrative data to see if the estimated effects are the same as those using scores from tests that researchers have given to the same sample of students. In TCA, for example, a replication researcher might bring in village-level data from another source to test for unexplored community effects. Retesting the original model using an entirely new data-set with similar variables drawn from the same population would be an example of MEA.

5. Other types of replication

Although this paper focuses on internal replication, there are other types of replication discussed in the literature, most notably external replication and implementation replication.

External replication is the study of the external validity of the research results by conducting the analysis on a different population. Hamermesh (2007) calls this scientific replication, where the same idea is examined with a different data-set drawn from a different population.

For impact evaluations of development programmes, external replication typically means a similar impact evaluation conducted on the same intervention implemented in a different population. One example is IPA's evaluations of the Consultative Group to Assist the Poor and the Ford Foundation's Ultra Poor Graduation programme.

In several different countries, the programme funds interventions designed to break the cycle of extreme poverty based on BRAC's Challenging the Frontiers of Poverty Reduction/Targeting the Ultra Poor model. Although the interventions in the different countries are all based on the same theory of change and general model, there are some minor variations in the services offered to accommodate local contexts. IPA is conducting RCTs of several of these interventions in order to test for both internal and external validity simultaneously.

Implementation replication is defined here as the evaluation of a new implementation of the same intervention, typically on the same population, but where the intervention is now being implemented at scale and/or by local institutions, not supervised by researchers. These replications are important for determining the feasibility and sustainability of piloted innovations. They can also be used to measure the cost effectiveness of programmes that local institutions have implemented at scale.

One recent example of implementation replication is the Bold *et al.* (2013) replication study of the Duflo *et al.* (2012) pilot educational intervention in Kenya. Duflo *et al*. tests whether a programme that Banerjee *et al*. (2007) and Muralidharan and Sundararaman (2011) study, which produces educational gains

for children by bringing an additional contract teacher into Indian schools, works in Kenya. They find significant educational gains for children taught by an NGO-contracted teacher versus those taught in the standard Kenyan education system.

To test the potential for national scale-up of this promising education intervention, Bold *et al*. (2013) conduct an implementation replication of Duflo *et al.* (2012) by copying the design of the programme but using state-contracted extra teachers instead of NGO contractors.

Although all the other factors of Bold *et al.*'s implementation remain constant, the impact evaluation finds that children taught by contract teachers from the Kenyan Ministry of Education do not demonstrate similar treatment effects as those in the Duflo *et al.* study. Local implementation of the contract teacher programme results in insignificant differences between the treated and control children. This example highlights the importance of implementation replication of proof of concept interventions to demonstrate the sustainability and generalisability of promising research to national programmes.

6. Conclusion

We are not the first to make the case that internal replication is needed to improve the quality and credibility of scientific or social scientific research. We are also not the first to discuss the various challenges in the research and publication industries that both create the need for, and yet hinder, the production and publication of replication research.

We have categorised these challenges with two objectives in mind. The first is to make a strong case for internal replication that is not predicated on catching bad researchers. Despite everything that has been written about replication to date, when we launched a replication programme at 3ie, many well-established researchers, as well as some funders, expressed grave concerns that such a programme would be seen as a witch hunt. In order to promote replication it is important to emphasise that replication studies are not accusations but rather responses to a faulty system and understandable mistakes, not to mention a central part of the scientific process.

The second objective is to provide more definition around the different analytical approaches that replication research can take. The literature to date assumes that internal replication includes pure replication and often some kind of additional analysis. There has been limited discussion of the kinds of analysis that can be done and how those approaches address the various challenges.

This paper defines three categories of replication analysis: pure replication, MEA and TCA. In practice, there is frequently overlap in the analysis of measurement, estimation and theory of change questions, but it is important to understand that the different questions address different sets of assumptions from an original study. As such, the results for the different questions may have distinct implications for the validity and credibility of the original findings for policy and programme recommendations.

The movement for evidence-based policymaking, whether in high-income or low- and middle-income countries, will only succeed if the evidence being provided for policymaking and programme design is credible, robust and truly elucidates the causal chain. Researchers generally have the best of intentions to produce such evidence but face very real challenges. Replication is not the whole solution, but it can play a key role in addressing the challenges, both *ex post* by testing the evidence and *ex ante* by changing incentives.

References

Abbott, A, 2007. Notes on replication. *Sociological Methods & Research,* 36(2), pp.210-19.

Acemoglu, D, Johnson, S, Robinson, JA and Yared, P, 2008. Income and democracy. *American Economic Review*, 98(3), pp.808–42.

Angrist, J 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), pp.3–30.

Auvert, B, Taljaard, D, Lagarde, E, Sobngwi-Tambekou, J, Sitta, R and Puren, A, 2005. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 trial. *PLoS Med*, 2(11), pp.1,112–22.

Bailey, RC, Moses, S, Parker, CB, Agot, K, Maclean, I, Krieger, JN, Williams, CFM, Campbell, RT and Ndinya-Achola, JO, 2007. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The Lancet*, 369(9562), pp.643–56.

Baker, D, 2013. In history of economic errors, Martin Feldstein deserves mention, [online] Available at: <http://www.cepr.net/index.php/blogs/beat-the-press/in-history-of-economic-errors-martin-feldstein-deserves-mention> [Accessed 12 February 2014].

Banerjee, A, Cole, S, Duflo, E and Linden, L, 2007. Remedying education: evidence from two randomized experiments in India. *Quarterly Journal of Economics*, 122(3), pp.1,235–64.

Basu, K and Foster, JE, 1998. On measuring literacy. *The Economic Journal*, 108(451), pp.1,733–49.

Bernanke, B, 2004. Editorial statement. *American Economic Review*, 94(1), p.404.

Berthon, P, Pitt, L, Ewing, M and Carr, CL, 2002. Potential research space in MIS: a framework for envisioning and evaluating research replication, extension, and generation. *Information Systems Research*, 13(4), pp.416–27.

Binmore, K and Shaked, A, 2010. Experimental economics: where next? *Journal of Economic Behavior and Organization*, 73(1), pp.87–100.

Bold, T, Kimenyi, M, Mwabu, G, Ng'ang'a, A and Sandefur, J, 2013. Scaling up what works: experimental evidence on external validity in Kenyan education. Working Paper 321. Center for Global Development.

Bueno de Mesquita, B, 2003. Getting firm on replication. *International Studies Perspectives*, 4(1), pp.98–100.

Burman, LE, Reed, WR and Alm, J, 2010. A call for replication studies. *Public Finance Review*, 38(6), pp.787–93.

Carter, MR and Barrett, CB, 2006. The economics of poverty traps and persistent poverty: an asset-based approach. *Journal of Development Studies*, 42(2), pp.178–99.

Cartwright, N, 1991. Replicability, reproducibility, and robustness: comments on Harry Collins. *History of Political Economy*, 23(1), pp.143–55.

Cattaneo, MD, Galiani, S, Gertler, PJ, Martinez, S and Titiunik, R, 2009. Housing, health, and happiness. *American Economic Journal: Economic Policy*, 1(1), pp.75–105.

Center for Effective Global Action, n.d. Housing, health and happiness: the impacts of cement flooring in Mexico. E2A Cement Floors Brief, [online] Available at: <http://cega.berkeley.edu/assets/cega_events/19/E2A_Cement_Floors_Brief.pdf> [Accessed 12 February 2014].

Cervellati, M, Jung, F, Sunde, U and Vischer, T, 2014. Income and democracy: comment. *American Economic Review*, 104(2), pp.707–19.

Chen, S and Ravallion, M, 2007. Absolute poverty measures for the developing world, 1981–2004. Policy Research Working Paper 4211. World Bank.

Collins, HM, 1984. When do scientists prefer to vary their experiments? *Studies in History and Philosophy of Science*, 15(2), pp.169–74.

Day, TE and Liebowitz, SJ, 1998. Mortgage lending to minorities: Where's the bias? *Economic Inquiry*, 36(1), pp.3–28.

Deaton, A, 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2), pp.424–55.

Dewald, WG, Thursby, JG and Anderson, RG, 1986. Replication in empirical economics: the journal of money, credit and banking project. *The American Economic Review*, 76(4), pp.587–603.

Donohue, JJ and Levitt, SD, 2001. The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2), pp.379–420.

Donohue, JJ and Levitt, SD, 2008. Measurement error, legalized abortion, and the decline in crime: a response to Foote and Goetz. *The Quarterly Journal of Economics*, 123, pp.425–40.

Duflo, E, Dupas, P and Kremer, M, 2012. School governance, teacher incentives, and pupil-teacher ratios: experimental evidence from Kenyan primary schools. Working Paper Series 17939. National Bureau of Economic Research.

Evanschitzky, H, Baumgarth, C, Hubbard, R and Armstrong, JS, 2007. Replication research's disturbing trend. *Journal of Business Research*, 60(4), pp.411–15.

Falk, R, 1998. Replication-a step in the right direction: commentary on Sohn. *Theory & Psychology*, 8(3), pp.313–21.

Fanelli, D, 2010. 'Positive' results increase down the hierarchy of the sciences. *PLoS One*, 5(4), pp.1–10.

Feldstein, M, 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy*, 82(5), pp.905–26.

Feldstein, M, 1982. Social security and private saving: reply. *Journal of Political Economy*, 90(3), pp.630–42.

Feldstein, M, 1996. Social security and saving: new time series evidence. *National Tax Journal*, 49(2), pp.151–64.

Folbre, N, 2013. Replicating research: austerity and beyond. *The New York Times*, [online] 22 April. Available at: <http://economix.blogs.nytimes.com/2013/04/22/replicating-research-austerity-and-beyond/?_php=true&_type=blogs&_r=0> [Accessed 12 February 2014].

Foote, CL and Goetz, CF, 2008. The impact of legalized abortion on crime: comment. *The Quarterly Journal of Economics,* 123(1), pp.407–23.

Freese, J, 2007. Overcoming objections to open-source social science. *Sociological Methods & Research*, 36(2), pp.220–26.

García, FM, 2013. Scientific progress in the absence of new data: a procedural replication of Ross (2006).

Gerber, A and Malhotra, N, 2008. Publication bias in empirical sociological research: do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1), pp.3–30.

Gherghina, S and Katsanidou, A, 2013. Data availability in political science journals. *European Political Science*, 12(3), pp.333–49.

Gleditsch, NP and Metelits, C, 2003. Replication in international relations journals: policies and practices. *International Studies Perspectives*, 4(1), pp.89–97.

Gleditsch, NP, Metelits, C and Strand, H, 2003. Posting your data: will you be scooped or will you be famous? *International Studies Perspectives*, 4(1), pp.89–97.

Goldberg, PK, 2013. American economic review. *American Economic Review: Papers and Proceedings*, 103(3), pp.701–12.

Gray, RH, Kigozi, G, Serwadda, D, Makumbi, F, Watya, S, Nalugoda, F, Kiwanuka, N, Moulton, LH, Chaudhary, MA, Chen, MZ, Sewankambo, NK, Wabwire-Mangen, F, Bacon, MC, Williams, CFM, Opendi, P, Reynolds, SJ, Laeyendecker, O, Quinn, TC and Wawer, MJ, 2007. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *The Lancet*, 369(9562), pp.657–66.

Hamermesh, DS, 2007. Viewpoint: replication in economics. *The Canadian Journal of Economics*, 40(3), pp.715–33.

Harrison, GW, 1998. Mortgage lending in Boston: a reconsideration of the evidence. *Economic Inquiry*, 36(1), pp.29–38.

Herndon, T, Ash, M and Pollin, R, 2013. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Working Paper Series 332. Political Economy Research Institute, University of Massachusetts, Amherst.

Herrnson, PS, 1995. Replication, verification, secondary analysis, and data collection in political science. *PS: Political Science and Politics*, 28(3), pp.452–55.

Horne, DK, 1994. Evaluating the role of race in mortgage lending. *FDIC Banking Review*, Spring-Summer, pp.1–15.

Horne, DK, 1997. Mortgage lending, race, and model specification. *Journal of Financial Services Research*, 11(1-2), pp.43–68.

Hubbard, R and Vetter, DE, 1992. The publication incidence of replications and critical commentary in economics. *American Economist*, 36(1), pp.29–34.

Humphreys, M, de la Sierra, RS and van der Windt, P, 2013. Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), pp.1–20.

IDEAS/RePEc, 2013. IDEAS/RePEc simple impact factors for journals, [online] Available at: <http://ideas.repec.org/top/top.journals.simple.html> [Accessed 12 February 2014].

Ioannidis, JP, 2005. Contradicted and initially stronger effects in highly cited clinical research. *Jama,* 294(2), pp.218–28.

Iversen, V and Palmer-Jones, R, forthcoming. TV, female empowerment and fertility decline in rural India.

Jensen, R and Oster, E, 2009. The power of TV: cable television and women's status in India. *Quarterly Journal of Economics,* 124(3), pp.1,057–94.

Jensen, R and Oster, E, 2012. Corrigendum, "The Power of TV." Available at: < http://faculty.chicagobooth.edu/emily.oster/papers/update.pdf> [Accessed 12 March 2014].

King, G, 1995. Replication, replication. *PS: Political Science and Politics,* 28(3), pp.444–52.

King, G, 2006. Publication, publication. *PS: Political Science and Politics*, 39(1), pp.119–25.

Leamer, EE, 1983. Let's take the con out of econometrics. *American Economic Review*, 73(1), pp.31–43.

Leamer, EE, 2010. Tantalus on the road to asymptopia. *The Journal of Economic Perspectives*, 24(2), pp.31–46.

Leimer, DR and Lesnoy, SD, 1982. Social security and private saving: new time-series evidence. *Journal of Political Economy*, 90(3), pp.606–29.

Lim, SS, Dandona, L, Hoisington, JA, James, SL, Hogan, MC and Gakidou, E, 2010. India's janani suraksha yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation. *The Lancet*, 375(9730), pp.2,009–23.

Lipset, SM, 1959. Some social requisites of democracy: economic development and political legitimacy. *American Political Science Review*, 53(01), pp.69–105.

Lipsey, MW and Wilson, DB, 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *The American Psychologist*, 48(12), pp.1,181–209.

Liverpool-Tasie, LSO and Winter-Nelson, A, 2011. Asset versus consumption poverty and poverty dynamics in rural Ethiopia. *Agricultural Economics*, 42(2), pp.221–33.

Mazumdar, S, Mills, A and Powell-Jackson, T, 2011. Financial incentives in health: new evidence from India's Janani Suraksha Yojana.

McCullough, B, 2009. Open access economics journals and the market for reproducible economic research. *Economic Analysis & Policy*, 39(1).

McCullough, B and McKitrick, R, 2009. Check the numbers: the case for due diligence in policy formation. Vancouver, BC: Fraser Institute.

Miguel, E and Kremer, M, 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), pp.159–217.

Mittelstaedt, RA and Zorn, TS, 1984. Econometric replication: lessons from the experimental sciences. *Quarterly Journal of Business and Economics*, 23(1), pp.9–15.

Munnell, AH, Browne, LE, McEneaney, J, and Tootell, GMB, 1992. Mortgage lending in Boston: Interpreting HMDA data. Working Papers 92-7. Federal Reserve Bank of Boston.

Munnell, AH, Tootell, GMB, Browne, LE and McEneaney, J, 1996. Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review*, 86(1), pp.25–53.

Muralidharan, K and Sundararaman, V, 2011. Teacher performance pay: experimental evidence from India. *Journal of Political Economy*, 119(1), pp.39–77.

Rampell, C, 2013. A history of oopsies in economic studies. *The New York Times*, [online] 17 April 2013. Available at: <http://economix.blogs.nytimes.com/2013/04/17/a-history-of-oopsies-in-economic-studies/> [Accessed 12 February 2014].

Reinhart, CM and Rogoff, KS, 2010. Growth in a time of debt. *s*, 100(2), pp.573–78.

Rokicki, S and Carvalho, N, forthcoming. Replication of 'India's Janani Suraksha Yojana'.

Rosenthal, R, 1979. The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, 86(3), pp.638–41.

Rothstein, HR, Sutton, AJ and Borenstein, M, 2005. Publication bias in meta-analysis: prevention, assessment and adjustments. Chichester, England: John Wiley & Sons.

Science's Sokal moment, 2013. *The Economist*, 5 October. p.85.

Sokal, A and Bricmont, J, 1998. Fashionable nonsense: postmodern intellectuals' abuse of science. New York : Picador USA.

Toncar, MF and Munch, JM, 2010. Meaningful replication: when is a replication no longer a replication? Rejoinder to Stella and Adam (2008). *Journal of Marketing Theory and Practice*, 18(1), pp.71–80.

Tweney, RD, 2004. Replication and the experimental ethnography of science. *Journal of Cognition and Culture*, 4(3-4), pp.731–58.

Unreliable research: trouble at the lab, 2013. *The Economist*, 19 October. pp.26–30.

Valentine, J, Biglan, A, Boruch, RF, Castro, FG, Collins, LM, Flay, BR, Kellam, S, Mościcki, EK and Schinke, SP, 2011. Replication in prevention science. *Prevention Science*, 12(2), pp.103–17.

Wible, JR, 1992. Fraud in science: an economic approach. *Philosophy of the Social Sciences*, 22(1), pp.5–27.

Wohlfarth, P, 2012. Replication in the narrow sense of Banzhaf/Walsh (2008). Replication Working Paper No. 02/2012. Georg-August University.

Yong, E, 2012. Replication studies: bad copy. *Nature*, May 17. pp.298–300.

Zakula, B, 2012. Narrow replication of Ashcraft (2005): are banks really special? Replication Working Paper No. 01/2012. Georg-August-University.

Appendix A: Journal replication policy survey results

a. Journals with replication support and data accessibility policy (21)

Economics journals (20):

American Economic Journal: Macroeconomics
American Economic Review
Brookings Papers on Economic Activity, Economic Studies Program
Econometrica
Econometrics Journal
Economic Journal
European Economic Review
International Economic Review
Journal of Applied Econometrics
Journal of Economic Perspectives
Journal of Environmental Economics and Management
Journal of Labor Economics
Journal of Law, Economics and Organization
Journal of Money, Credit and Banking
Journal of Political Economy
Journal of the European Economic Association
Review of Economic Dynamics
Review of Economic Studies
The Review of Economics and Statistics
World Bank Economic Review

Development journals (1):

American Economic Journal: Applied Economics

b. Journals with only data accessibility policy (1)

Economics journals (1):

Journal of Human Resources

c. Journals with a suggested or informal replication policy (10)

Economics journals (9):

Carnegie-Rochester Conference Series on Public Policy
Economic Policy
Journal of Business & Economic Statistics
Journal of Economic Growth
Journal of Economic Theory
Journal of Financial Economics
Journal of Financial Intermediation
Journal of International Economics
World Bank Research Observer

Development journals (1):

Journal of Development Studies

d. Journals with no replication or data accessibility policy (17)

Economics journals (9):

Journal of Accounting and Economics
Journal of Econometrics
Journal of Finance
Journal of LACEA Economia
Journal of Law and Economics
Journal of Public Economics
Journal of Risk and Uncertainty
Quarterly Review
Review of Financial Studies

Development journals (8):

Development Policy Review
Economic Development and Cultural Change
Economics & Human Biology
European Journal of Development Research
Journal of Development Effectiveness
Journal of International Development
Oxford Development Studies
World Development

e. Replication policy not applicable or no answer to inquiries (15)

Economics journals (12):

Economic Policy Review
Experimental Economics
Journal of Development Economics
Journal of Economic Literature
Journal of Economic Surveys
Journal of International Money and Finance
Journal of Monetary Economics
Oxford Bulletin of Economics and Statistics
Proceedings (Board of Governors of the Federal Reserve System (U.S.))
Proceedings (Federal Reserve Bank of Cleveland)
Proceedings (Federal Reserve Bank of San Francisco)
The Quarterly Journal of Economics

Development journals (3):

Development and Change
Journal of Economic Development
Review of Development Economics

Replication Paper Series

International Initiative for Impact Evaluation
1625 Massachusetts Ave., NW
Suite 450
Washington, DC 20036
USA

3ie@3ieimpact.org
Tel: +1 202 629 3939

www.3ieimpact.org