Routledge
Taylor & Francis Group

Journal of Development Effectiveness

ISSN: 1943-9342 (Print) 1943-9407 (Online) Journal homepage: http://www.tandfonline.com/loi/rjde20

The growth of impact evaluation for international development: how much have we learned?

Drew B. Cameron, Anjini Mishra & Annette N. Brown

To cite this article: Drew B. Cameron, Anjini Mishra & Annette N. Brown (2016) The growth of impact evaluation for international development: how much have we learned?, Journal of Development Effectiveness, 8:1, 1-21, DOI: 10.1080/19439342.2015.1034156

To link to this article: http://dx.doi.org/10.1080/19439342.2015.1034156

Journal of Development Effectiveness, 2016
Vol. 8, No. 1, 1–21, http://dx.doi.org/10.1080/19439342.2015.1034156

Routledge
Taylor & Francis Group

The growth of impact evaluation for international development: how much have we learned?

Drew B. Cameron[a]*, Anjini Mishra[b] and Annette N. Brown[a]

[a]*International Initiative for Impact Evaluation, 1625 Massachusetts Ave NW, Suite 450, Washington, DC 20036, USA;* [b]*Independent Consultant, 1625 Massachusetts Ave NW, Suite 450, Washington, DC 20036, USA*

This article examines the content of a web-based repository of published impact evaluations of international development interventions. To populate this repository, we conducted a systematic search and screening process. We find that of the 2259 studies published from 1981 to 2012, annual publication increased dramatically after 2008. Most studies are on health, education, social protection and agriculture and are concentrated in South Asia, East Africa, South and Central America and Southeast Asia. There are statistically significant differences in time between end line data collection and publishing by the publication type, and institutional affiliation of authors has shifted towards countries in North America and Europe.

Keywords: impact evaluation; international development; development effectiveness; evidence database; evidence-based policy

1. Introduction

In recent years, the international development community has seen an increased focus on the commissioning and use of impact evaluation to inform programming. By 'impact evaluation', we refer to counterfactual-based programme evaluation that attempts to attribute specific outcomes to programmatic activities by dealing with the problem of selection bias (for a complete discussion, see White 2010). As the evaluation gap working group argued in its 2006 report 'Will we ever learn? Improving Lives through Impact Evaluation', building this evidence helps 'to improve the effectiveness of domestic spending and development assistance by bringing vital knowledge into the service of policymaking and program design' (CGD 2006, 1). Among the recommendations of the working group nearly a decade ago was the creation of a 'comprehensive database of impact studies' with the aim to facilitate access to rigorous evidence (36).

To meet this recommendation, the International Initiative for impact evaluation (3ie) launched a database of impact evaluations in 2009, indexing more than 700 studies by the end of 2012. From 2013 to 2014, the database underwent a retrofit to systematically collect and index all available published impact evaluation evidence. The new impact evaluation repository (IER) was inspired by the work of the Campbell Collaboration's Sociological, Psychological, Educational, and Criminological Trials Register (C2-SPECTR).[1] C2-SPECTR was conceived to be a comprehensive online international database of randomised controlled trials (RCTs) to help policy makers and researchers identify the most rigorous evidence of what works across various disciplines (Turner et al.

*Corresponding author. Email: dcameron@3ieimpact.orgr

2003). The IER similarly aims to help researchers and policy makers identify experimental and quasi-experimental evaluations of international development programmes, projects and policies by summarising and cataloguing all IE evidence in one common web-based portal.

This article outlines the systematic search and screening process used to populate the new IER and presents the initial findings. To our knowledge, no other database or research endeavour has attempted to systematically index all published impact evaluation evidence of international development programmes. As such, this article presents the first complete overview of the state of this popular research methodology for international development. This article describes the evidence base from 2259 impact evaluations carried out from 1981 through 2012 in 145 low- and middle-income countries and territories.[2] The majority of these studies were identified through a systematic search and screening of content from key databases, websites and research libraries that took place between January 2013 and June 2014. Unlike a systematic review or meta-analysis, we did not code outcome variables or effect sizes and thus do not examine the results of these evaluations. However, the systematic collection of this evidence base does allow a powerful starting point for future synthesis studies (see for example Vivalt 2015).

This article proceeds in Section 2 by providing details of the systematic search and screening methods undertaken from 2013 to 2014, and coding details recorded for each study. In Section 3, we present the results of our analysis including the timing and source of study publication, sectors evaluated and evaluation methods used, geographic locations where evaluations have taken place, the average time taken between data collection and publication of impact evaluations across a number of publication types and the regional distribution of authors' institutional affiliations. In Section 4, we present a discussion of these findings. Section 5 outlines some of the limitations of this research project, and Section 6 provides conclusions.

2. Methods

In order to locate impact evaluation evidence across all international development sectors, we developed a search and screening protocol for 45 different online academic databases, organisation websites, search engines, journal collections and research libraries.[3] This search protocol is meant to be implemented semi-annually. We also received impact evaluations submitted by website users as well as those identified through other literature reviews and snowball sampling techniques (as described in Section 2.2). All identified studies were screened according to pre-defined selection criteria, then coded and uploaded to the repository. It is important to reiterate that studies were not assessed or coded based on their quality.

2.1. Search strategy and inclusion criteria

The first systematic search took place between January and June 2013. We searched major academic databases in health, economics, public policy and the social sciences including those provided by platforms such as Ovid, EbscoHost and ProQuest. Online libraries and websites were also included from select research organisations and academic institutions. For a complete list of online resources searched refer to Appendix 1.[4]

We identified potential impact evaluation research from these sources through the development of individual systematic search strategies for each resource, taking into account the unique features and limitations of each search platform. We chose a set of

subject terms and free-text keywords in three principal thematic areas: geography (drawn from the World Bank's classification of low- and middle-income countries), programme evaluation terms (including thesaurus and free text terms like 'program* evaluation,' 'impact evaluation' and 'randomised controlled trial'), and impact evaluation keywords including 'impact', 'evaluation', 'assessment', 'effect*', 'random*', 'trial', 'intervention', so forth and their permutations.[5] When possible, we included exclusionary terms with the Boolean operator 'NOT', such as 'NOT systematic review' and 'NOT meta-analysis'.

2.2. Other study collection methods

Between June and July 2014, after the inclusion of all studies found during the systematic search and screening process of the previous 18 months, we initiated a crowdsourcing effort on the 3ie website asking visitors to submit any impact evaluations not already in the repository. Studies submitted which met our inclusion criteria but were not present in the repository were eligible for a $10 gift certificate. The contest received around 90 impact evaluation entries from 27 people. These studies were screened using the repository inclusion criteria (see Mishra and Cameron 2013). Sixty-three of these studies met all the inclusion criteria and were added to the repository. Of these, 32 were published prior to the initiation of our search and screening protocol.[6]

Other methods of collecting impact evaluations are also regularly in use. Before the search and screening protocol outlined in this article went into effect in early 2013, the IER held 704 studies. These were collected through a number of processes. One method, 'snowballing', was used to locate studies by searching the cited references within impact evaluations. Studies are also regularly submitted to the repository by authors who wish to have them included. Other studies are identified during the process of systematic searching in other 3ie knowledge products including systematic reviews and gap maps. Finally, studies commissioned by 3ie are kept in a separate 3ie-funded studies database. Once these studies have been published as final reports, they are automatically included in the Impact evaluation repository. All studies identified through these other methods must meet the inclusion criteria (see Mishra and Cameron 2013).

2.3. Screening process

We screened all studies for the following inclusion criteria: (1) the full text of the study must be available in English, (2) the study must be published, (3) the study must evaluate an intervention that took place in a low- or middle-income country, (4) the study must evaluate at least one specific policy, programme or intervention, (5) the study must use a valid impact evaluation method and (6) the study must evaluate programme effectiveness.[7] When in doubt regarding any of these criteria, we erred on the side of inclusion, favouring type-two over type-one errors. Furthermore, we endeavoured to refrain from making value judgements about the quality of research reviewed for the IER and attempted to adhere to a static interpretation of our selection criteria.

We conducted the first screening process from June 2013 to April 2014 (see Figure 1). A total of 30,061 studies were rejected in the initial title screening stage.[8] In stage 2, title and abstract information for 19,427 records was reviewed between two screeners (Drew Cameron and Anjini Mishra). Rejected studies were verified by each screener. In stage 3, each of the remaining 8570 studies were assigned to four teams of two screeners each who independently reviewed the full text of each study using the inclusion criteria in Mishra and Cameron (2013). Each screener generated a score of 'Yes', 'Unclear' or 'No'. Scores

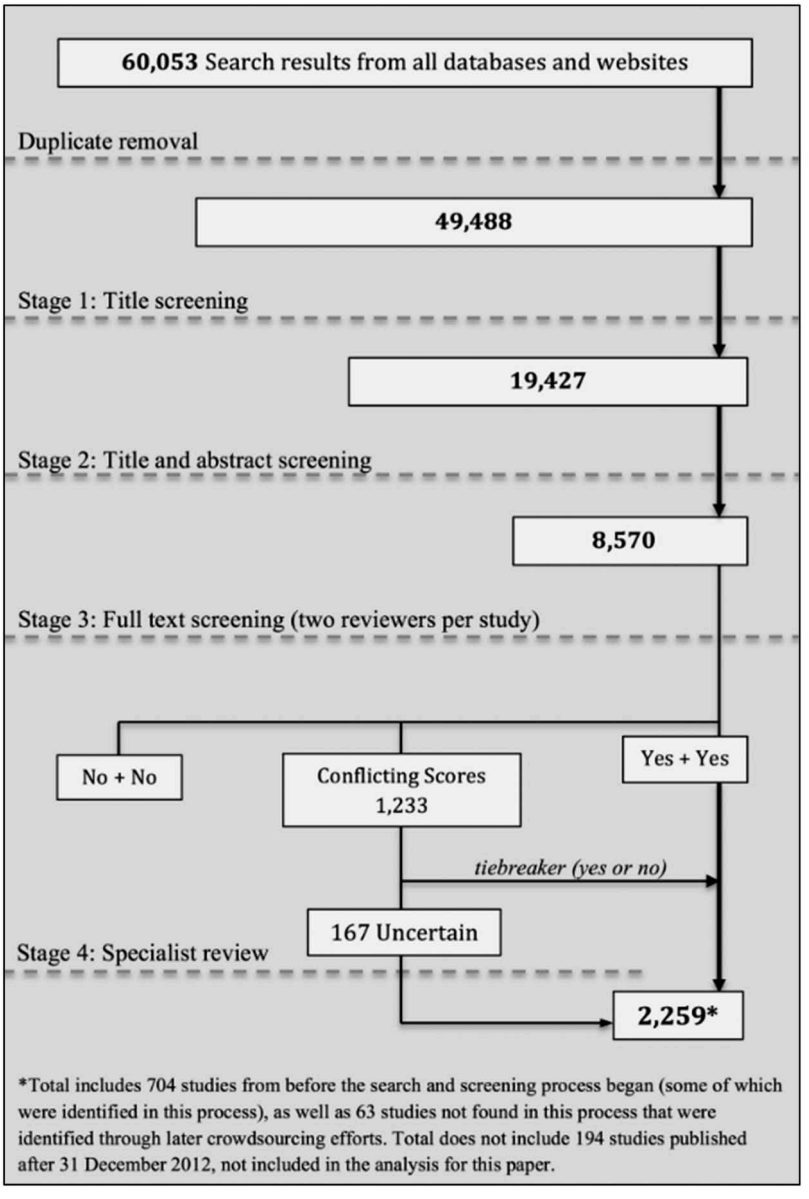

Figure 1. Screening flow chart (May 2013–March 2014).
Source: Author constructed.

were compared as illustrated in Figure 1. Studies with conflicting scores (1233 in total) were reviewed by a third screener. If a tiebreaking score was not reached, the study was sent for a fourth stage of 'specialist review'.[9]

Once included in the IER, studies were coded for the following information: study title, author(s), publication date, countr(y/ies), region(s), sector(s), subsector(s), evaluation method(s) (difference-in-differences, instrumental variable estimation, randomised controlled trial, propensity score matching and other matching methods and regression

discontinuity design), publication type (journal article, working paper, report, book or book chapter), selected sub-group(s) (gender, conflict-afflicted, differently-abled, elderly, ethnic minorities, indigenous groups, migrant workers, orphans and vulnerable children and refugees) and bibliographic details. Coding details were verified by two screeners and then uploaded.[10]

2.4. *Additional subgroup analysis*

In order to engage in additional analysis outlined in Sections 3.4 and 3.5, we collected a random sample of 147 studies from the IER. Of these, 130 were published prior to 2013 (5.8% of the 2259 studies published before 2013) and were used for analysis. The mean comparisons between the random sample and the overall dataset are described in Table 1.[11] For each randomly selected study, the final date of study data collection was

Table 1. Mean comparisons, pre-2013 studies vs. random sample.

		Full dataset (pre-2013)	Random sample (pre-2013)
Total studies		2259	130
	% pre-2013 studies	.	5.8%
Journal article		1740	100
	% column total	77.0%	76.9%
Working paper		426	26
	% column total	18.9%	20.0%
Report		87	4
	% column total	3.9%	3.1%
Book (or book chapter)		6	0
	% column total	0.3%	0.0%
RCT only		1394	85
	% column total	61.7%	65.4%
Mixed RCT & quasi-experiment		104	6
	% column total	4.6%	4.6%
Quasi-experiment only		761	39
	% column total	33.7%	30.0%
Health journal		1224	65
	% column total	54.2%	50.0%
Social science journal		528	35
	% column total	23.4%	26.9%
Bank or international lending agency		216	11
	% column total	9.6%	8.5%
Government agency		18	3
	% column total	0.8%	2.3%
Research institute or university		273	16
	% column total	12.1%	12.3%
Americas		595	42
	% column total	26.3%	32.3%
Asia		909	56
	% column total	40.2%	43.1%
Africa		716	33
	% column total	31.7%	25.4%
Europe		69	1
	% column total	3.1%	0.8%

recorded when the full text was available. This information was used to determine the number of years between the final date of data collection and publication. The full text was available for 126/130 studies published prior to 2013. Of these, 13 studies did not report the date of end line data collection, leaving 113 studies for end line data collection analysis in Section 3.4.

We also recorded the institutional affiliations by country for every study author listed in each of the 130 randomly selected evaluations. This information was used to conduct analysis on the regional location of study authors in Section 3.5. Data on author affiliations were available for 129/130 randomly selected studies published prior to 2013. The study with no author information (a report) was coded as having one single author from the country where the report was published. Authors with multiple institutional affiliations in both high-income and low- or middle-income countries were coded as having their primary affiliation in the high-income country reported, unless primary affiliation otherwise stated. All data analysis for this article was conducted using Stata SE version 12.

3. Findings

The following section examines this evidence from a number of vantage points. First, we outline patterns of impact evaluation publication over time, noting a precipitous rise in published impact evaluations from 2000 to 2012. In Section 3.2, we examine countries and regions where impact evaluations have taken place taking special note of where evidence is most abundant and most lacking. Section 3.3 examines some of the sectors where impact evaluation evidence is most abundant. Section 3.4 provides an overview of the time typically taken to produce impact evaluation evidence between sectors and types of evaluation. Finally, Section 3.5 examines the institutional affiliations of impact evaluation authors.

3.1. Impact evaluation publications over time

A first look at the body of published impact evaluation evidence shows that the publication of this kind of evidence took some time to take hold. Beginning in the mid-1990s, impact evaluation evidence was being published at a steady rate. By the turn of the century, impact evaluation publications were increasing at a more rapid rate, continuing to the year 2012 (Figure 2). As Bill Savedoff points out, this increase in evidence publication does correspond to a few seminal events like the first published evaluations of the conditional cash transfer programme Progresa in the late 1990s, the creation of institutions like the Abdul Latif Jameel Poverty Action Lab (J-PAL), the World Bank's Development Impact Evaluation Initiative (DIME) and the Strategic Impact Evaluation Fund (SIEF) at the World Bank.[12]

Out of 132 studies published prior to 2000, 107 (81.1%) were published in health journals, 12 (9.1%) were published in social science journals, banks and other international lending agencies published 7 studies (5.3%) and 6 (4.5%) were published by research institutions, universities or non-governmental organisations (NGOs). Only towards the end of the 1990s did impact evaluations begin to appear in non-health publications. Prior to 1996, only 6 (of 36) studies had been published outside of health sciences journals.

After the year 2000, impact evaluations began to appear more frequently in social science journals, from international agencies and governments (usually as reports), and from research institutions and universities, as illustrated in Figure 3. From 2000 to 2004, health journal publications still represented 60% of the published impact evaluation evidence. But, starting

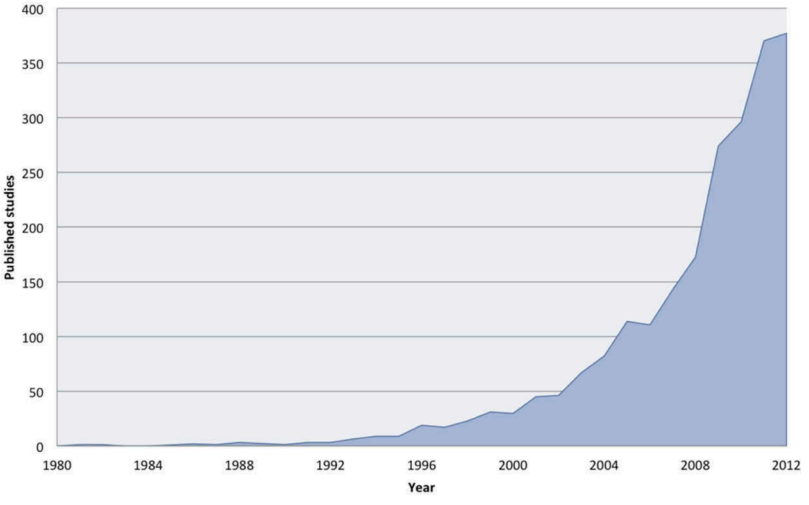

Figure 2. Impact evaluations published per year (1981–2012).

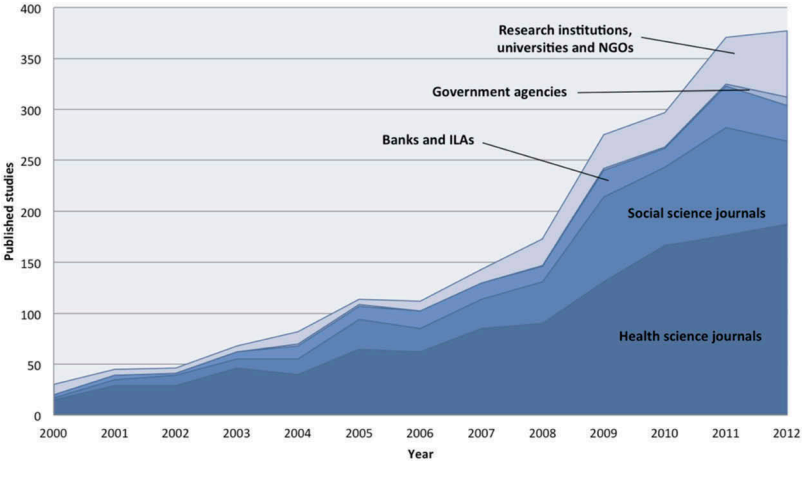

Figure 3. Impact evaluations published by source (2000–2012).
Note: Because data collection took place starting in early 2013, studies in the IER published after 2012 do not adequately represent the total body of impact evaluation evidence after 2012.

in 2004, publications from non-health sources began to increase, representing 47% of the total share of published evidence from 2005 to 2009. During this five-year period, the most dramatic single-year rise in overall publications took place between 2008 and 2009 when studies published per year increased by 63% (from 173 to 274 studies).

Some trends could be explained by the publication process in different disciplines. In Figure 3, a 37% increase in publications from research institutions and universities can be seen from 2011 to 2012. A majority (84%) of publications from these research institutions and universities in 2012 were working papers. This large increase in working paper publications could eventually lead to an increase in the number of social science journal publications. Similar patterns are not visible in health journal publications.

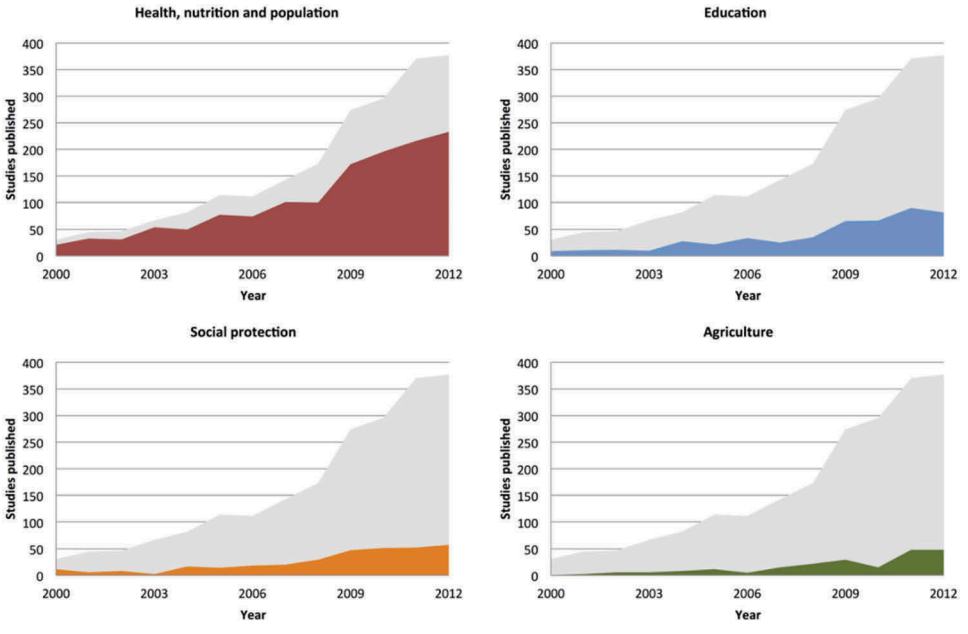

Figure 4. Impact evaluations published by major sector (2000–2012).

3.2. *In what sectors have impact evaluations taken place?*

By the turn of the century, impact evaluations were becoming more popular in sectors outside of health, nutrition and population. Figure 4 shows the share of all impact evaluations between the four largest sectors. Health, nutrition and population still dominated the total number of impact evaluations after the year 2000, though other sectors see modest increases as well. Many studies overlap between multiple categories, so relative comparisons can be misleading. However, we can tell that in the 10 years after 2000, 54% (584/1085) of studies dealt with subject matters outside of health, and in the first 3 years of the current decade (2010–2019), the share has risen to 58% (606/1043).

When we examine the trends across all sectors more closely (see Table 2), a steady increase in impact evaluations conducted in most sectors is apparent over the last 12 years. The total number of evaluations of health, population and nutrition programmes continues to increase year to year, however, these seem to be met with equally steady increases in the other three major sectors (agriculture, social protection and education). Those sectors where the rigorous evidence base continues to stagnate include economic policy, energy, transportation and urban development.

Table 3 shows studies in each sector by use of at least one of the five impact evaluation methods employed, difference-in-differences (DD), instrumental variable estimation (IV), randomised controlled trials (RCT), regression discontinuity design (RDD) and propensity score matching or other matching methods (PSM or OMM). Studies may employ multiple methods and be counted in multiple categories.

Randomised controlled trials dominate the evaluation literature in just a few sectors. Evaluations of health nutrition and population interventions predominately use randomisation (83%) to evaluate programme effectiveness. This rate is even higher when isolating only those studies published in health journals (92.8%). Other sectors where evaluations

Table 2. Impact evaluations published per year by sector (2000–2012).

Sector	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	% of IER pre-2013
Agriculture and rural development	1	2	6	6	8	12	5	15	22	29	15	48	48	9.7%
Economic policy	0	1	0	0	0	0	0	1	0	4	6	2	2	0.7%
Education	9	11	12	10	28	22	34	25	35	66	67	90	82	23.1%
Energy	0	0	0	0	1	1	0	2	0	2	1	6	1	0.6%
Environment and disaster management	0	0	1	1	0	1	3	8	7	8	11	18	17	3.4%
Finance	0	1	2	1	3	6	4	9	15	14	17	28	17	5.5%
Health, nutrition and population	21	33	31	54	50	78	74	101	100	172	196	216	233	64.9%
Information and communications technology	0	0	0	0	1	1	9	3	3	6	8	14	18	2.8%
Private sector development	0	1	2	3	3	5	5	8	5	12	22	26	21	5.1%
Public sector management	0	2	0	3	2	4	3	5	6	14	6	13	13	3.3%
Social protection	12	6	8	2	17	14	18	20	29	47	51	52	57	15.1%
Transportation	0	0	0	0	1	1	1	1	0	0	2	2	1	0.4%
Urban development	0	0	1	0	1	2	0	1	2	4	4	1	3	0.8%
Water and sanitation	0	1	2	5	4	7	5	5	8	12	13	12	13	4.2%
Total studies	30	45	46	67	82	114	111	143	173	274	296	370	377	.

Note: Column totals do not reflect the sum of rows within each column as each study may be catalogued in multiple sector categories; final column of table includes proportion of studies within IER from 1981 to 2012.

Table 3. Total number of studies by sector and the impact evaluation method used.

Sector		DD	IV	RCT	RDD	PSM or OMM	*Total*
Agriculture and rural development		70	46	45	0	96	*219*
	% sector	*32.0%*	*21.0%*	*20.5%*	*0.0%*	*43.8%*	
Economic policy		7	5	0	1	6	*16*
	% sector	*43.8%*	*31.3%*	*0.0%*	*6.3%*	*37.5%*	
Education		109	57	304	33	92	*521*
	% sector	*20.9%*	*10.9%*	*58.3%*	*6.3%*	*17.7%*	
Energy		5	5	0	0	8	*14*
	% sector	*35.7%*	*35.7%*	*0.0%*	*0.0%*	*57.1%*	
Environment and disaster management		21	17	23	2	26	*76*
	% sector	*27.6%*	*22.4%*	*30.3%*	*2.6%*	*34.2%*	
Finance		37	22	51	0	30	*124*
	% sector	*29.8%*	*17.7%*	*41.1%*	*0.0%*	*24.2%*	
Health, nutrition and population		140	65	1228	11	96	*1476*
	% sector	*9.5%*	*4.4%*	*83.2%*	*0.7%*	*6.5%*	
Information and communications technology		16	2	42	2	13	*63*
	% sector	*25.4%*	*3.2%*	*66.7%*	*3.2%*	*20.6%*	
Private sector development		42	22	30	8	39	*115*
	% sector	*36.5%*	*19.1%*	*26.1%*	*7.0%*	*33.9%*	
Public sector management		23	17	33	4	12	*74*
	% sector	*31.1%*	*23.0%*	*44.6%*	*5.4%*	*16.2%*	
Social protection		107	42	139	21	117	*341*
	% sector	*31.4%*	*12.3%*	*40.8%*	*6.2%*	*34.3%*	
Transportation		5	1	2	1	2	*9*
	% sector	*55.6%*	*11.1%*	*22.2%*	*11.1%*	*22.2%*	
Urban development		7	3	3	1	7	*19*
	% sector	*36.8%*	*15.8%*	*15.8%*	*5.3%*	*36.8%*	
Water and sanitation		13	5	68	0	16	*94*
	% sector	*13.8%*	*5.3%*	*72.3%*	*0.0%*	*17.0%*	
Total		377	187	1499	54	363	*2259*
	% total	*16.7%*	*8.3%*	*66.4%*	*2.4%*	*16.1%*	

Notes: DD = difference-in-differences, IV = instrumental variable estimation, RCT = randomised controlled trial, RDD = regression discontinuity design, PSM or OMM = propensity score matching or other matching method; methods used are not exclusive. For example, a study may employ multiple methods and/or be double-counted in multiple sector categories; health, nutrition and population sector designation does not imply that studies were published in health journals; chart includes all studies from 1981 to 2012.

predominately use randomisation to evaluate programme effectiveness are education (60%), information and communications technology (68%) and water and sanitation services (69%).

Impact evaluations in most other sectors employ quasi-experimental methods quite frequently. In the case of agriculture and rural development interventions, for example, propensity score matching is a frequently used method. In evaluation of transportation programmes, difference-in-differences is used most frequently. Finally, instrumental variable estimation seems to be used frequently in evaluation of interventions in energy, economic policy, agriculture and environment and disaster management.

3.3. *Where are impact evaluations conducted?*

The vast majority of low- and middle-income country impact evaluations are concentrated in countries with the largest populations. In fact, the 10 most populous low- and middle-income countries in the world (CIA 2014)[13] account for over 41.3% (933) of all development impact evaluations we collected. Just three of these countries (India, China and Mexico) make up almost a quarter (22.5%) of all impact evaluation evidence published before 2013. However, when controlling for population density (Table 4), the most populous nations rank towards the bottom of the list in terms of impact evaluation studies published per million people. Mexico stands out as having the highest density of studies among the 10 most populous developing nations. Interestingly, half (50.1%) of impact evaluations in Mexico published prior to 2013 are related to conditional cash transfers. Meanwhile, the most densely studied countries (that is countries with the highest study per population rates) are mostly island nations with small populations.[14]

Most impact evaluation evidence comes from studies conducted in a single country (96.2%), as opposed to multi-country studies (3.8%). According to Table 5, impact evaluation evidence seems to be concentrated in South Asia (21.9%) and Eastern Africa (19.0%). Meanwhile, South America (14.7%), Central America (10.7%) and Southeast Asia (10.4%) comprise a lower (but still substantial) share of the overall evidence base. The regions with the greatest relative lack of impact evaluation evidence seem to be Central and Northern Africa, central and West Asia (including the Middle East), Eastern Europe and Oceania (see also Figure 5). Currently, there are only 14 countries for which there is no impact evaluation evidence in any sector.[15]

Table 4. IE density of 10 most populous low- and middle-income countries (1981–2012).

Country	Population	Number of impact evaluations	IE count rank	IE studies per 1 m pop.	IE density rank
China	1,355,692,576	141	3	0.1	135
India	1,236,344,631	215	1	0.17	126
Indonesia	253,609,643	64	9	0.25	120
Brazil	202,656,788	96	7	0.47	100
Pakistan	196,174,380	64	9	0.33	113
Nigeria	177,155,754	28	28	0.16	128
Bangladesh	166,280,712	110	6	0.66	84
Russia	142,470,272	14	41	0.01	137
Mexico	120,286,655	152	2	1.26	55
Philippines	107,668,231	49	21	0.46	104

Note: Population information from the CIA World Factbook 2014.

Table 5. Number of impact evaluations published by region per decade (1981–2012).

		1980–1989	1990–1999	2000–2009	2010–2012	Total	% of total
Americas		**4**	**27**	**323**	**241**	***595***	***26.3%***
	Central America (including Mexico)	*2*	*15*	*137*	*87*	*241*	*10.7%*
	South America	*2*	*11*	*176*	*143*	*332*	*14.7%*
	Caribbean	*0*	*4*	*20*	*13*	*37*	*1.6%*
Africa		**3**	**51**	**311**	**352**	***717***	***31.7%***
	Northern Africa	*0*	*2*	*9*	*12*	*23*	*1.0%*
	Southern Africa	*1*	*5*	*64*	*56*	*126*	*5.6%*
	Western Africa	*2*	*10*	*62*	*73*	*147*	*6.5%*
	Middle Africa	*0*	*1*	*6*	*12*	*19*	*0.8%*
	Eastern Africa	*0*	*33*	*181*	*217*	*431*	*19.0%*
Europe		**0**	**6**	**34**	**29**	***69***	***3.1%***
	Eastern Europe	*0*	*4*	*25*	*22*	*51*	*2.3%*
	Northern Europe	*0*	*0*	*5*	*5*	*10*	*0.4%*
	Southern Europe	*0*	*2*	*6*	*5*	*13*	*0.6%*
Asia		**4**	**38**	**431**	**436**	***909***	***40.2%***
	Central Asia	*0*	*0*	*3*	*3*	*6*	*0.3%*
	Eastern Asia	*0*	*5*	*63*	*78*	*146*	*6.5%*
	South Asia	*2*	*22*	*230*	*240*	*494*	*21.9%*
	Southeast Asia	*2*	*11*	*122*	*99*	*234*	*10.4%*
	Western Asia	*0*	*0*	*21*	*22*	*43*	*1.9%*
Oceania		**0**	**0**	**9**	**7**	***16***	***0.7%***
	Melanesia	*0*	*0*	*2*	*3*	*5*	*0.2%*
	Micronesia	*0*	*0*	*1*	*0*	*1*	*0.0%*
	Polynesia	*0*	*0*	*5*	*5*	*10*	*0.4%*

Note: Region designations as defined by the United Nations.

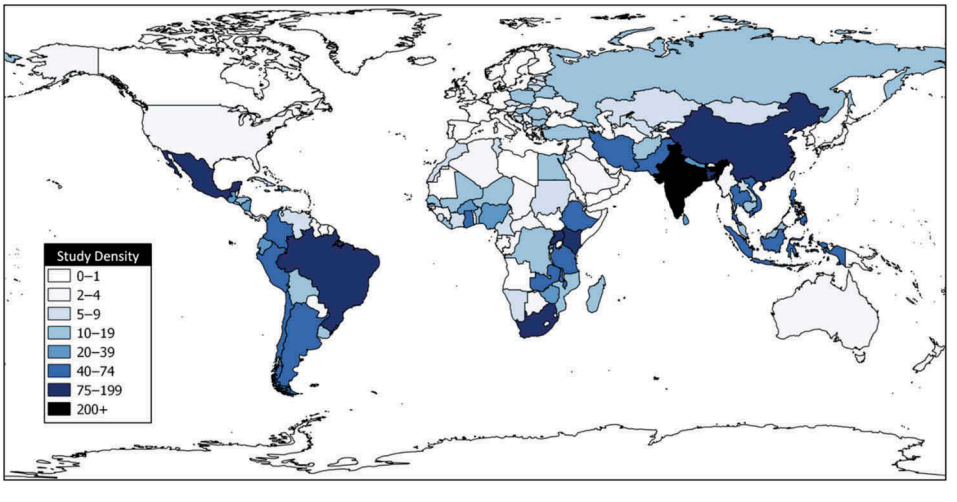

Figure 5. Heat map of low- and middle-income country impact evaluations (1981–2012).
Note: Map generated using shape files from Natural Earth in Quantum GIS version 2.2.0. Free vector and raster map data @ naturalearthdata.com.

3.4. How timely is published evidence?

Table 6 shows the results of our test of the mean difference in the number of years between end line data collection and year of publication for four different publication types: journal articles, working papers, reports and books or book chapters. The majority of studies published prior to 2013 are journal articles (77%), which take an average of 4.69 years from end line data collection for findings to be published. Working papers have a shorter time lag by about one year on average, while reports produce evidence that is published relatively quickly (1.00 years on average). We tested the difference in means using a one-way ANOVA (variance of means) test and found that the difference in time from end line data collection to publication is highly significant (at 99%) across the different publication types.

Table 7 attempts to disentangle the publication sources to determine if there are substantial differences in time taken to publish. It seems that among journal articles, social science publications take almost twice as long to publish as impact evaluations in health journals (6.18 vs. 3.75 years, respectively). Banks and international lending agencies (3.55 years), as well as universities and other research institutions (3.54 years) are comparable to health journals. Government agencies, meanwhile, lead the way in producing expeditious evaluations (at 1.00 years). Again, a one-way ANOVA test of the variance of means shows that these differences between different publication sources are statistically significant at the 99% level.

Table 8 shows the difference in time between end line data collection and publication across different evaluation methodologies. The initial estimates show no statistically significant difference in the time to publication between randomised controlled trials, mixed designs and quasi-experimental only studies. We also ran additional *t*-tests for each evaluation methodology, as well as for each method individually and found that only for studies using instrumental variable estimation was there a statistically significant difference between mean years from end line data collection to publication (7.1 years, vs. approximately 4 years for non IV papers). However, as illustrated in Table 9, the sample of instrumental variable articles is very small (only 10 impact evaluations in our random sample). When we tested additional variations such as difference in the mean years for articles using experimental vs. quasi-experimental methods published in health vs. social science journals, we also found no statistically significant differences.

3.5. Author institutional affiliations

Figure 6 illustrates the differences in author institutional affiliations when grouped by region. From our random sample of 130 impact evaluations, nearly half of all authors (49.7%) list institutional affiliations within countries in North America, Western or Northern Europe. The other half is made up of authors with affiliations in Latin America and the Caribbean (15.2%), South Asia (12.8%), sub-Saharan Africa (8.2%), Southeast Asia and Oceania (6.6%), East Asia (4.4%) and the Middle East and North Africa (2.3%). The proportion of first author affiliations is skewed slightly more towards North American and European country institutions (59.2%), while the remainder is made up by authors with affiliations in Latin America (13.8%), South Asia (9.2%), Southeast Asia (6.2%), sub-Saharan Africa (4.6%), East Asia (3.8%) and the Middle East and North Africa (3.1%).

However, the regional distribution of author affiliations does seem to have changed over time. In Figure 7, we examine the regional distribution of authorship between two time periods, 1981–2008 and 2009–2012. As illustrated previously in Figure 2, the most substantial year-to-year increase in published impact evaluation evidence took place

Table 6. Mean years from end line data collection to publication by publication type (pre 2013).

		Journal article	Working paper	Report	Book or book chapter	Total	ANOVA (*f* stat)
All studies (pre-2013)		1740	426	87	6	*2259*	
	% all studies	*77.0%*	*18.9%*	*3.9%*	*0.3%*		
Random sample[a]		86	24	11[a]	5[a]	*126*	
	% publication type	*4.9%*	*5.6%*	*12.6%*	*83.3%*	*5.6%*	
Mean years from end line to publication		**4.69**	**3.63**	**1**	**4.8**	***4.17***	***4.98******
	Standard deviation	(3.189)	(3.321)	(1.483)	(2.387)	*(3.232)*	

Notes: ***Significant at the 99% confidence level; [a] 'Report' and 'Book or Book Chapter' observations were oversampled in order to capture a large enough sample for data analysis; this accounted for 5 books and 8 additional reports published prior to 2013 (total $n = 126$) randomly selected from the remaining un-sampled studies in the IER. The ANOVA test reported here included this set of 13 additional studies. Without adding these 13 studies, mean differences by publication type are significant at 95% confidence (mean values: journal article (same), working paper (same), reports (3) – mean = 0.33, standard deviation = 0.577; books (0); ANOVA *f* stat = 3.48**).

Table 7. Average years from data collection to publication by source of publication (pre 2013).

		Health journal	Social science journal	Bank or ILA	Government agency	University or research institute	ANOVA (*f* stat)
All studies pre-2013		1224	528	216	18	273	
	% all studies	*54.2%*	*23.4%*	*9.6%*	*0.8%*	*12.1%*	
Random sample (n = 113)		53	33	11	3	13	
	% publication source	*4.3%*	*6.3%*	*5.1%*	*16.7%*	*4.8%*	
Mean years from end line to publication		**3.75**	**6.18**	**3.55**	**1**	**3.54**	**4.78*** **
	Standard deviation	(2.295)	(3.836)	(3.328)	(1.732)	(3.526)	

Notes: ***Significant at the 99% confidence level; oversampled 'report' and 'book or book chapter' studies from Table 6 were not included in this analysis, but we did run a robustness check using these additional 13 studies (2 from banks and international lending agencies, 2 from government agencies and 9 from universities and research institutes). Mean years from end line to publication were within 0.20 of the values reported here: Banks and ILAs (3.38, sd. = 3.228), government agencies (0.08, sd. = 1.304) and universities and research institutions (3.36, sd. = 3.140). The ANOVA test was still significant at 99% confidence (*f* stat = 6.19).

Table 8. Average years from end line data collection to publication by the evaluation methodology.

		Experimental	Mixed	Quasi-experimental	ANOVA (*f stat*)
All studies pre-2013		1394	104	761	
	% all studies	*61.7%*	*4.6%*	*33.7%*	
Random sample (n = 113)		70	4	39	
	% method	*5.0%*	*3.8%*	*5.1%*	
Mean years from end line to publication		**3.9**	**6.5**	**4.92**	**2.19**
	Standard deviation	(2.869)	(3.674)	(3.801)	

Note: 'Experimental' includes all studies that use randomisation to allocate treatment and control conditions, and do not use quasi-experimental methods; 'Quasi-experimental' includes all studies that do not use an 'Experimental' methodology; 'Mixed' includes all studies that use both randomisation and at least one quasi-experimental method (difference-in-differences, instrumental variable, propensity score matching or other matching method or regression discontinuity design).

Table 9. *t*-Test mean years from end line data to publication by method.

		Yes	No	*p-Value*
Difference-in-differences		3.67	4.45	*0.389*
	Standard error	(0.532)	(0.344)	
	n	15	98	
Instrumental variable estimation		7.4	4.05	*0.002***
	Standard error	(1.74)	(0.278)	
	n	10	103	
Propensity score matching or other matching method		4.1	4.4	*0.712*
	Standard error	(0.624)	(0.349)	
	n	20	93	
Randomised controlled trial		4.04	4.92	*0.172*
	Standard error	(0.342)	(0.602)	
	n	74	39	
Regression discontinuity design		4.67	4.34	*0.863*
	Standard error	(1.764)	(0.312)	
	n	3	110	

Notes: ***Significant at the 99% confidence level; methods are not mutually exclusive; n = 113 for each *t*-test.

between 2008 and 2009 (a 63% increase). The majority (58.3%) of impact evaluations published until 2012 were printed over the four-year period from 2009 to 2012. We thus separately consider these two distinct periods in the publication of impact evaluation evidence.

First, we find that the total number of authors increased substantially from the pre-2009 period (251) to 2009–2012 (321). When we examine author distributions between these two periods it is clear that from 2009 to 2012 the total share of authors with institutional affiliations in North America and Western and Northern Europe increased over 10 percentage points (pp), from 43.8% to 54.2%. In real terms, the total number of authors from non-North American and European institutions also rose slightly after 2008, from 141 authors to 147 authors. However, these regions were stagnant or even decreased in terms of the total share of author affiliations (Latin America (+0.1%), South Asia

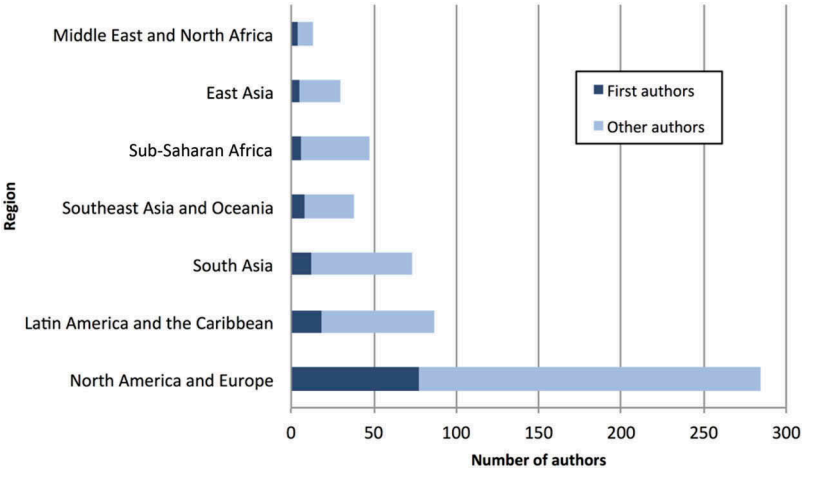

Figure 6. Random sample of author institutional affiliation by region.

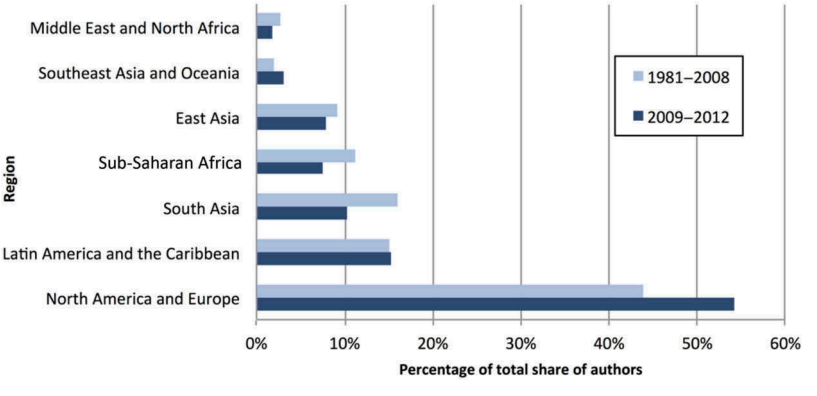

Figure 7. All authors' institutional affiliation by region.

(−5.7%), Southeast Asia and Oceania (+1.1%), sub-Saharan Africa (−3.7%), East Asia (−1.4%) and Middle East and North Africa (−0.9%). We see similar trends when examining only first author institutional affiliations, where the share in North America and Europe increased 9.0 pp between the two periods (from 53.8% to 62.8%), the share of those in East Asia increased 6.4 pp (from 0.0% to 6.4%) and shares for all five other regions decreased (from between 1.2 and 7.1 pp).

4. Discussion

This study confirms many commonly held beliefs about impact evaluation research for international development, and throws others into question. In the last 14 years, there has been a precipitous rise in the publication of these evaluations across many sectors, regions and disciplines. Specifically, large publication increases are apparent between 2004 and 2005 and most dramatically between 2008 and 2009. Publications in health sciences journals have dominated the bulk of available impact evaluation evidence throughout all periods, though in

recent years there seems to be a proliferation of studies published outside of traditional academic journals by research institutions, universities and NGOs.

Impact evaluation evidence of international development interventions is mostly concentrated in just a few geographic regions. Most notably, South Asia, East Africa and much of Latin America hold the bulk of the available evidence. By contrast, rigorous counterfactual-based evaluations are much less abundant in the Caribbean, francophone Africa and Central Asia. Evidence is practically non-existent in Oceania, the Middle East, North Africa and Central Africa.

The sectors where we have seen an increase in the production of impact evaluation evidence are health, nutrition and population, education, agriculture and rural development and social protection. However, there has not been a large increase in the production of impact evaluation evidence in all sectors. For example, very little impact evaluation evidence exists of interventions in transportation, energy, economic policy and urban development.

The remaining sector areas have experienced only a modest increase in impact evaluation evidence in recent years. Evaluations of development interventions in the realms of public sector management, environment and disaster management, private sector development and water and sanitation services have all been published more frequently since around 2009, but still nowhere near the extent of studies in the four major sector areas.

In terms of time taken to publish results, we measured the difference in years from end line data collection to the publication date for a random sample of published impact evaluations. We found a substantial and statistically significant difference between types of publications. Journal articles in social sciences are easily the most time-intensive form of published impact evaluation evidence (over 6 years between end line data collection and publication on average). Similar studies in health journals take about half the time (3.75 years), and are on par with publications from banks and international lending agencies (3.55 years) and universities and research institutions (3.54 years). However, the most expeditious evidence clearly comes from impact evaluations commissioned from government agencies (1.00 years). It is important to note that this study does not make any claims about the quality of these impact evaluations outside of inclusion criteria including the study identification strategy.

Through our analysis we do not find any statistically significant differences between years from end line data collection to publication according to evaluation methodology. This suggests that the methodology plays very little role in the speed of impact evaluation evidence production and publication. Indeed, there is no difference between exclusively experimental and quasi experimental studies. The only impact evaluation method with a significant difference in time taken to publish is instrumental variable estimation (7.40 years for IV vs. 4.05 years for others). However, the sample of studies using instrumental variable estimation in our analysis is quite small ($n = 10$), and we have little confidence that this result will hold up to further scrutiny.

Finally, it is remarkable to note that among a random sample of 130 impact evaluations of interventions in low- and middle-income countries, 50% of all authors, and 59.2% of first authors are from countries in North America and Western Europe. It is also remarkable that in the four years after 2008 (during what seems to be the initial heyday of impact evaluation publication) there was a substantial increase in both the number and the total share of impact evaluation authors with institutional affiliations in North America and Western and Northern Europe. This share is even larger among first authors, though we recognise that first authorship does not necessarily equate to primary authorship. This

finding may seem to fly in the face of the intentions of many major funders of impact evaluation research who hope to increase low- and middle-income country capacity to produce impact evaluation evidence. However, in real terms, we do find that the total number of authors has continued to increase steadily across all regions.

5. Limitations

This is the first iteration of the impact evaluation repository under the new systematic search and screening protocol. As such, a number of shortcomings are worth note. Since data collection was initiated in early 2013, studies published after 2012 should not have been captured by the IER search strategy. As such, studies published in 2013 and 2014 are not included in the analysis for this article, though we do note that trends have not differed much for the available evidence from 2013 to 2014. Additionally, screeners were encouraged to err on the side of inclusion when approving studies for the repository. This may have led to a slightly increased (though probably negligible) number of studies on the margins of the inclusion criteria.

Upon an initial review of the databases used for this search, we determined that several databases and websites were not included in the first systematic search. As such, this collection may underrepresent studies from public policy resources, studies from the social protection literature, studies available in some health-specific resources (such as Cochrane), and recently launched institutional resources (like the World Bank's new Impact Evaluation Working Paper Series). Additionally, the crowdsourcing effort from June to July 2014 revealed a number of other new web-based resources that will be added to the protocol in future. These missing resources also highlight that the database may have under-sampled reports and grey literature such as working papers.

The current body of evidence reviewed in this article does not include published studies for which the full text version is unavailable in English. In future, we also plan to include non-English titles, as well as to reform and expand the current region, sector and subsector categories, review and potentially update the current list of evaluation methods (to differentiate natural experiments and other methodologies) and to include coding options for studies that employ cost-effectiveness or cost-benefit analysis.

It is also worth reiteration that neither this study, nor the impact evaluation repository itself, attempts to assess the quality of any included impact evaluations. Our inclusion criteria very explicitly dissuaded screeners from including or excluding content based on a judgement of a study's quality. Instead, screeners were instructed to determine (yes or no) whether at least one of the identification strategies (noted earlier) was used appropriately.

6. Conclusions

The impact evaluation repository is the first resource of its kind to systematically collect counterfactual-based evaluation evidence of developing country interventions. Since the need for a comprehensive database was first identified nearly a decade ago, impact evaluation evidence has undergone dramatic growth in a number of sectors using a variety of methodologies. For the first time, a clear picture of this growth has emerged, allowing researchers and policymakers to better examine evaluation trends across disciplines, production processes from region to region, and to identify important gaps in the literature. The growth of impact evaluation evidenced in this article does suggest that we are finally starting to realise the need for more rigorous research in international development. However, we must caution that just because a great deal of impact

evaluation evidence now exists, does not necessarily mean that it has catalysed greater learning among researchers and policymakers.

Acknowledgements

Special thanks are due to the IER screening and summary writing team, including Adrian Scutaru, Amy Holter, Celeste Visser, Flor Calvo, Hisham Esper, Igor Louboff, Isaac Hurr, Laura Carpenter, Michael Broache, Nikita Salagonkar and Sarah Oberst. Thanks are due also to specialist reviewers including Anna Heard, Benjamin Wood, Eric Djimeu and Howard White. Thanks are due to Alexis Shenfil Smart for formatting assistance. Finally, the authors wish to thank peer reviewers at the *Journal of Development Effectiveness* for their thoughtful feedback and suggestions for future research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. This centralised 'register' was designed to help researchers overcome some of the most 'formidable challenge[s to] efficiently locating the highest possible number of RCTs through a single web-based platform that would be routinely updated through 'retrospective and prospective surveillance systems' (Turner et al. 2003, 204).
2. The IER currently contains over 2500 impact evaluations. However, more than 200 of these have been published since the beginning of 2013, after our systematic search and screening had begun. As such, our results only contain an accurate cross-section of the literature published before our initial data collection began.
3. Mishra, Anjini and Drew Cameron. 2013. 'IER Search and Screening Protocol.' *International Initiative for Impact Evaluation*. Washington, DC, pp. 159. Available Online: http://www.3ieimpact.org/media/filer_public/2014/05/23/3ie_repository_protocol.pdf
4. For a detailed explanation of the resources utilised, the specific search terms used, and the choice behind the selection of each database, website, and library refer to Mishra and Cameron (2013).
5. See Mishra and Cameron (2013) for a detailed explanation of search strategies for each resource.
6. Of the 32 studies missed in the search and screening protocol, 17 were journal articles, 9 were reports and 6 were working papers. The 17 journal articles should have been captured in our process and future rounds of searching have been modified to account for databases where these studies were indexed. The 15 working papers and reports were from sources unknown to us at the time of searching. These resources have also been added to the search and screening protocol.
7. Details on each of the six inclusion criteria can be found in Mishra and Cameron (2013).
8. Initially 32,037 records were rejected. These rejected records were then verified by a second screener who selected a random sample of 5% of rejected records from each day (batch) of screening in stage 1 (from a total of 40 batches of studies) and conducted a more thorough screening of the title and abstract. In total, 1976 previously rejected records were selected for further screening and added to the 17,451 studies previously accepted in stage 1.
9. Specialists included Eric Djimeu, Anna Heard, Benjamin Wood, and Howard White.
10. Coding also includes 300 word summaries of study methodology, and main findings after all other details have been recorded on the IER website. The process of summary writing is ongoing.
11. Reports and books were oversampled in order to obtain a larger number of studies for subgroup analysis in Table 6. This oversampling is explained in detail the note for that table.
12. See: http://www.cgdev.org/blog/hot-topic-cool-heads-impact-evaluation-debated-cgd-3ie-conference

13. China, India, Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, Russia, Mexico and the Philippines.
14. The top 10 most densely studied countries, in order (studies per 1 million population in parentheses), are Tonga (6), Palau (1), Samoa (4), American Samoa (1), Sao Tome and Principe (1), Nicaragua (33), Gambia (11), Jamaica (13), the Solomon Islands (3) and Malta (1). In Nicaragua, 57.6% of impact evaluations examine conditional cash transfer programmes.
15. Democratic People's Republic of Korea, Dominica, Federated States of Micronesia, Grenada, Kiribati, Kosovo, Maldives, Marshall Islands, Saint Lucia, Seychelles, Somalia (and Somaliland), St. Vincent and the Grenadines, Tuvalu and Republic of Yemen.

References

CGD. 2006. *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, DC: Center for Global Development, Evaluation Gap Working Group.

CIA. 2014. *CIA world factbook*. Accessed October 14. https://www.cia.gov/library/publications/the-world-factbook/

Mishra, A., and D. Cameron. 2013. "Impact Evaluation Repository Search and Screening Protocol." International Initiative for Impact Evaluation, Washington, DC. Accessed October 20, 2014. http://www.3ieimpact.org/media/filer_public/2014/05/23/3ie_repository_protocol.pdf

Turner, H., R. Boruch, A. Petrosino, J. Lavenberg, D. De Moya, and H. Rothstein. 2003. "Populating an International Web-Based Randomized Trials Register in the Social, Behavioral, Criminological, and Education Sciences." *The ANNALS of the American Academy of Political and Social Science* 589: 203–223. doi:10.1177/0002716203256840.

Vivalt, E. 2015. "How Much Can We Generalize from Impact Evaluations?" Working Paper. Accessed March 10. http://evavivalt.com/wp-content/uploads/2014/11/Vivalt-JMP-11.06.14.pdf

White, H. 2010. "A Contribution to Current Debates in Impact Evaluation." *Evaluation* 16 (2): 153–164. doi:10.1177/1356389010361562.

Appendix 1. Database, website and library resources utilised in the search process

Africa Wide Information (Ebsco), CAB Abstracts (Ebsco), EconLit (Ebsco), SocINDEX (Ebsco), Academic Search Complete (Ebsco), Embase (Ovid), Medline (Ovid), PsycINFO (Ovid), British Library of Development Studies, EppiCentre, Education Resources Information Center (ERIC), IDEAS RePEc, JOLIS, Popline, Sage Journals, Science Direct, Wiley Online Library, Trip Database, International Bibliography of the Social Sciences (IBSS), Social Sciences Resource Network (SSRN), New York Academy of Medicine, Web of Knowledge: Web of Science, Google Scholar, BREAD Working Papers, Centre for Global Development, International Food Policy Research Institute (IFPRI), Jameel Abdul Latif Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA), the World Bank, World Bank Development Impact Evaluation (DIME) Initiative, IE2 Impact Evaluations in Education (World Bank), Rural Education Action Project (REAP), Inter-American Development Bank, DFID Resources, Overseas Development Institute, Chronic Poverty Research Centre (CPRC), Governance and Social Development Resource Centre, NBER Working Papers, Center for Effective Global Action (CEGA), Asian Development Bank, Poverty and Economic Policy Research Network, USAID Development Experience Clearinghouse, OECD Development Assistance Committee Evaluation Resource Center (DEReC), African Development Bank, and Millennium Challenge Corporation (MCC).