Broadening horizons in impact evaluation for water, sanitation and hygiene planning: recycling and reinterpreting evidence

Hugh James Waddington

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

of the
University of London

July 2021

Department of Disease Control

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funding details: funding was gratefully received from the Campbell Collaboration, Japan International Cooperation Agency, the Millennium Challenge Corporation, UK Medical Research Council, the United States Agency for International Development, the Water Supply and Sanitation Collaborative Council, and through my staff position supported by the International Initiative for Impact Evaluation.

Research group affiliation(s): Environmental Health Group
London International Development Centre

I, Hugh James Waddington, confirm that the work presented in this Thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the Thesis.

Abstract

To meet universal Sustainable Development Goal targets, decision-makers need evidence about the effectiveness policy and programmes. Impact evaluations aim to provide that evidence, by quantifying the magnitude of changes in outcomes caused by WASH interventions in particular contexts for particular groups. However, there are concerns about the findings of single studies like randomised controlled trials (RCTs) and non-randomised studies (NRS), due to biases inherent in each approach. The best way to inform decisions is to use evidence from a variety of methodologies and contexts.

An evidence census shows that, while the quantity and quality of WASH impact evaluations has increased, there are important ethical concerns about relevance, reporting and representativeness. Drawing on the census, a critical appraisal tool was developed to evaluate consistently biases in RCTs and NRS. The tool was piloted in systematic reviews of internal and external replications on international development topics. The results of systematic reviews and meta-analyses that applied the tool in external replications were analysed. The findings showed that NRS with relatively low risk-of-bias produced the same pooled effects on average as RCTs (standardised mean difference (SMD)=0.00; 95% confidence interval (CI)=-0.06, 0.06), but NRS with high risk-of-bias over-estimated effects (SMD=0.17; 95% CI=0.07, 0.28). A systematic review of internal replication studies also found well-designed NRS produced effects that were statistically indistinguishable from RCTs (mean squared error=0.00).

Lack of access to and use of safe water, sanitation and hygiene (WASH) are thought to kill 300,000 children annually. RCTs are often considered the best causal evidence, but they cannot usually assess mortality due to power and ethical reasons. Existing systematic reviews assume diarrhoea morbidity is closely correlated with mortality. Meta-analysis of mortality impacts from the evidence census found 15 percent reduction in the odds of all-cause mortality in childhood, and 50 percent reduction in odds of diarrhoea mortality. WASH interventions reduce more deaths when they include hygiene and total sanitation.

Table of Contents

List of figures

List of tables

List of boxes

Acknowledgements

This Thesis is about recycling. If it has a primary objective, it is to make better use of what is already known, using the data already collected and state-of-the-art methods in causal inference and evidence synthesis, to provide evidence for policy and programmes to improve people's lives. If it has a goal, it is to contribute to the movement for more development resources to get to where they should be going – policymakers, practitioners, programme evaluators and, ultimately, programme participants in low-income settings.

I came to the Department of Disease Control as a part-time PhD student shortly after a 're-review' was published (Loevinsohn et al., 2015) of a systematic review I had led on the effectiveness and sustainability of water, sanitation and hygiene (WASH) programmes in combatting diarrhoea infection in childhood. Our review had extolled the virtues of incorporating behaviour change in impact evaluations and systematic reviews, by explicitly collecting outcomes data along the causal pathway and using theory to interpret findings (one of the first 'theory-based systematic reviews'), but we could have better incorporated theory and evidence on water-washed disease transmission. After applying to study with Professor Sandy Cairncross (Order of British Excellence), who was quick to point out that the departmental researchers were 'not just a bunch of diarrhoea heads', I had envisaged that the Thesis would focus on socioeconomic outcomes of importance for poor people – especially time savings, income and safety from improved water supply and sanitation. It does veer into socioeconomic territory, particularly in Chapter 5 which concerns impact evaluation and synthesis research on development topics outside of the WASH sector. However, my interests have also come full circle, as can be seen in Chapter 6, which focuses on diarrhoea mortality.

Many people helped and encouraged me, especially in the final two years of the project. My supervisor, Sandy, and co-supervisor, Dr Edoardo Masset gave superb guidance and support, and endured long delays on receipts of drafts. Professor Howard White, an early career mentor and employer, provided helpful direction on possible publication routes for some of the chapters. Others helped inform different aspects of the Thesis. I have been especially

fortunate to work closely with Dr Jorge Garcia Hombrados (University of Madrid), Professor Peter Tugwell and Dr Vivian Welch (University of Ottawa). I am grateful to my co-authors, listed under Contributions below. John Eyers (Hoop Cottage) helped design and run the searches on which the reviews in Chapters 3-6 are based. Professor Ruth Stewart (University of Johannesburg and ACE) gave helpful inputs for Chapter 3. Drs Adam Biran, Katie Greenland and Belen Torondel-Lopez enabled my teaching on environmental health courses at LSHTM, providing inspiration for Chapter 6. Dr Dean Spears (University of Texas at Austin and Research Institute for Compassionate Economics, r.i.c.e.) kindly provided estimates incorporated into the analysis in Chapter 6. Professors Simon Cousens (LSHTM) and Philip Davies (University of Oxford) supported the Thesis progression as members of the upgrading committee, which was chaired by Dr Jeroen Ensink. I also thank the examiners, Dr Anne Peasey (University College London) and Professor Paul Hunter (UEA).

Profound gratitude goes to Shubh Sharma, who helped with the maintenance of my mental hygiene, and was patient and supportive during the writing-up period. In the last few months of the project, during the first lockdown due to the global pandemic, Shubh and my dear Mum, Dr Sheila Waddington, helped with data collection for Chapter 3. My amazing sister, Clare Waddington, inspired an early interest in international development and global ethics. My sisters, nephews and nieces provided a lot of encouragement and emotional sustenance: Dr Kate Hagger, Daniel and God-daughter Eleanor; Megan Keirnan and Peter; Anna Lidgate, Alex and Evelyn. Our father sadly died a month before I enrolled in the degree. Being originally a town planner, and subsequently a sociologist, he was not a fan of economics (or, more likely, ‘economist supremacists’). I would argue with him while studying BSc in Economics and MA in Development Economics that, despite the faults in the way economics is often taught and misused politically, it is just another philosophy of science that can be applied to planning to make the world the better place.

The notion of ‘Buddhist Economics’ was helpful during the last year of the Thesis. “The Buddhist point of view takes the function of work to be at least threefold: to give a [person] a chance to utilise and develop [their] faculties; to enable [them] to overcome [their] egocentredness by joining with other people

in a common task; and to bring forth the goods and services needed for a becoming existence" (E.F. Schumacher, 1973, p.39). It is the opposite of "the new order… [where] every figure is trying to survive by concentrating on [their] own immediate need and survival… And faced with such reductionism, human intelligence is reduced to greed" (adapted from an open letter by the Subcomandante Marcos of the Zapatista National Liberation Army, by Richard Harold Kirk in Electronic Eye, Neurometrik, Alphaphone Recordings).

I dedicate this Thesis to Jeroen Ensink, a firm believer in collaboration who helped steer the focus to a more manageable topic and structure.

Contributions

The following contributions were made to this Thesis.

Chapter 3: Electronic literature searches designed and run by John Eyers. Additional electronic and hand search sifting and data collection by Hannah Chirgwin, Yashaswini PrasannaKumar, and Dua Zehra.

Chapter 4: indicative risk of bias approach developed with contributions from Jorge Garcia Hombrados (JH). Additional critical appraisals in Section 4.6 by Juliette Finetti, JH and Jennifer Stevenson.

Chapter 5: Electronic literature searches designed and run by John Eyers. Additional searches for studies in Section 5.3 by Paul Fenton Villar (PFV). Additional data extraction in Section 5.3 by Chris Coffey and PFV.

Chapter 6: Sarah Bick critically appraised and calculated effect sizes for a random selection of studies.

Chapter 1 The value of impact evaluation and evidence synthesis for global poverty reduction efforts

1.1 Introduction

> "The state of the public health of a community is determined at any particular time by the interaction of many diverse influences. Some of these influences are good some are bad; some are known, others unknown… The task of the public health service is to take cognisance of all these influences; to assess the effects of them; to foster the good ones, and to attempt to eliminate the bad ones."
>
> M'Gonigle and Kirby (1937, pp.19-20)

Water, sanitation and hygiene (WASH) are human rights that underpin basic needs. Most fundamentally, WASH affects the likelihood of survival beyond early childhood, and determines whether basic needs for human life – nutrition, excretion and safety – and higher order needs – like dignity, productivity, and happiness – are met (Maslow, 1943). Yet, according to the World Health Organization (WHO) and United Nations Children's Fund (UNICEF) Joint Monitoring Programme (JMP) for Water Supply and Sanitation, 2 billion people do not have safe, readily available water at home, and 4.5 billion lack access to safely managed sanitation services (WHO/UNICEF, 2019). How can this be, when the technologies and resources exist to provide everyone with safely managed WASH, when improved WASH provides the foundation for combating communicable diseases like diarrhoea which is endemic in low-income communities, killing millions every year, as well as for blocking infectious disease transmission in epidemics, such as the coronavirus disease 2019 (COVID-19) pandemic (Howard et al., 2020)?

At least part of the reason is due to competing priorities among decision-makers. To meet universal targets as defined by the Sustainable Development Goals (SDGs), decision-makers need access to evidence on what are the most effective ways to provide access to and promote use of WASH services, in

particular contexts, and for specific groups, particularly those who are the hardest to reach like remote populations and the most disadvantaged.

M'Gonigle and Kirby's (1937) evaluation of slum upgrading in 1920s Stockton-on-Tees, England – one of the first impact evaluations of a large-scale public health intervention – quoted above, indicated the great interest and challenges in attributing changes in quality of life to environmental health improvements.[1] Impact evaluations are attribution studies that aim to quantify the magnitude of effect of WASH provision or use on outcomes like child survival. There has been rapid growth in impact evaluations, especially randomised controlled trials (RCTs), owing to the influx of resources from major funders like the Gates Foundation. RCTs are not always feasible or ethical, but there is a debate about whether non-randomised studies (NRS) are able to produce unbiased estimates of effect. In addition, single studies, of whatever design, provide information specific to the context in which they are conducted, and may not be communicated in a way that is relevant or accessible for decision-making. Hence, there has been a simultaneous rise in evidence synthesis, particularly systematic reviews, which aim to provide critically appraised findings about generalisability of the evidence to aid decision-making.

This Thesis draws these different strands together on the effects of WASH policy and programmes in low- and middle-income countries (L&MICs), impact evaluation using randomised and non-randomised appoaches, and evidence synthesis. This first chapter introduces the Thesis topic, covering WASH sector interventions (Section 1.2), the consequences of limited access to and use of WASH (section 1.3), and presents a causal framework linking interventions and outcomes (Section 1.4). Section 1.5 discusses approaches to evaluating causal relationships using randomised and non-randomised evaluation. Section 1.6 discusses bias in design and implementation of evaluation studies and evidence synthesis methods that aim to overcome bias. The final section overviews the Thesis chapters.

[1] Quotes from M'Gonigle and Kirby (1937) are used throughout this chapter to highlight the many points raised in that classic study which remain relevant for evaluating environmental health impacts in low-income contexts.

1.2 Water, sanitation and hygiene interventions

> "The physical condition of our population is now less unsatisfactory than it was 30 years ago but, whatever degree of improvement has taken place should not be allowed to blind us to the present state of affairs which, as has been shown, still remains unsatisfactory."
>
> M'Gonigle and Kirby (1937, pp.179-180).

The quality of water supply, sanitation and hygiene facilities – that is, the extent to which they are likely to provide drinking water of sufficient quantities for basic needs, enable hygienic handwashing and food preparation, and safe removal of excrement from the human environment – is dependent on the types of water, sanitation and hygiene technology available. These have been articulated into ladders providing the indicators against which global progress is measured (Table 1.1).[2]

There has been broad consensus on the need for universal access to improved WASH since the 1977 United Nations (UN) Water Conference at Mar del Plata and subsequent International Drinking Water Supply and Sanitation Decade of the 1980s. The goal of that Decade, ratified by the Conference, was to provide adequate access to safe water and hygienic latrines to the population of the world by 1990 (Cairncross et al., 1980: xi). In 1990, the Convention on the Rights of the Child recognised the "right of the child to the enjoyment of the highest attainable standard of health… through the provision of… clean drinking water, taking into consideration the dangers and risks of environmental pollution" (Article 24, p.57; cited in Jolly, 2004, p.274). In the intervening decades, the UN has coordinated global indicators for improved access to and use of WASH facilities, and the targets set to measure their achievement.

[2] There are also intermediate steps on the sanitation ladder not listed in Table 1.1. For example, where there is no fixed place of sanitation but some attempt to remove faeces from exposure to others such as 'cat sanitation' (Waterkeyn and Cairncross, 2005).

Table 1.1 Ladders of WASH technology improvements

	Drinking water	*Sanitation*	*Hygiene*
Improved facilities: safely managed	Improved facilities that: • are accessible on premises, and • provide water when needed, and • provide water free from contamination.	Improved facilities where waste products are either: • treated and disposed in situ, or • temporarily stored and then emptied and transported to off-site treatment centre, or • transported through sewer with wastewater and treated off-site.	Undefined.
Improved facilities: basic	Improved sources that require less than 30 minutes round-trip to collect (including queueing time). These include piped supplies: • tap water in the dwelling, yard, or plot • public standposts/ pipes. And non-piped supplies: • boreholes/ tube wells • protected wells and springs • rainwater • packaged water, including bottled water and sachet water • delivered water, including trucks and small carts.	Improved facilities provided at the household level. These include networked sanitation: • flush and pour flush toilets connected to sewers. And on-site sanitation: • flush or pour flush toilets connected to septic tanks or pits • pit latrines with slabs • composting toilets, including twin pit latrines and container-based systems.	Fixed or mobile handwashing facilities with soap and water: • handwashing facilities defined as a sink with tap water, buckets with taps, tippy-taps, and jugs or basins designated for handwashing • soap includes bar soap, liquid soap, powder detergent, and soapy water.
Limited facilities	Improved sources of the above types requiring more than 30 minutes to collect including queueing time.	Improved facilities of the above types shared by two or more households.	Handwashing facilities without soap and water (e.g., ash, soil, sand or other handwashing agent).
Unimproved facilities	Non-piped supplies: • unprotected wells and springs.	On-site sanitation or shared facilities of the following types: • pit latrines without slabs • hanging latrines • bucket latrines.	Undefined.
No facilities	Surface water (e.g., drinking water directly from a river, pond, canal or stream).	Open defecation (disposal of human faeces in open spaces or with solid waste).	No handwashing facility on premises.

Sources: WHO/UNICEF (2017, 2019); https://washdata.org/monitoring.

The Millennium Declaration in 2000 included a water goal, and, following a declaration at the World Summit on Sustainable Development at Johannesburg in 2002, a sanitation goal was added (Jolly, 2004). The resulting Millennium Development Goal (MDG) 7 drinking water and sanitation targets were to halve (from 1990 levels) the proportion of people without sustainable access to safe drinking water and basic sanitation by 2015. The water indicator was later further defined as access to water from an improved source within 1 kilometre of the household. This is roughly the time taken for a 30-minute round-trip to collect water in the absence of queueing, which has been demonstrated as the time up to which basic needs for water supply can be reasonably met (White et al., 1972; Cairncross and Feachem, 2018). There are circumstances where it is likely that more than 30 minutes will be needed for 1 kilometre roundtrips, such as mountainous or sandy terrain, or in water scarce regions where people may spend more time queuing at the water collection point than travelling to it (Dar and Khan, 2011).[3] It is worth noting that the apparatus has been in place to monitor progress on water collection times at national (rural and urban) level in most countries at least since the Demographic and Health Surveys (DHSs) included a question on the time taken to "go there, fetch water, and come back" in Phase II in 1988-1993 (Institute for Resource Development/Macro International, 1990). JMP has since defined improved drinking water as 'basic' when it requires less than 30 minutes round-trip to collect (see also Table 1.1).

The Agenda for Sustainable Development set new global targets for 2030, enshrined in the SDGs.[4] The SDGs are more ambitious than the MDGs, aiming to "ensure the availability and sustainable management of water and sanitation for all" by 2030 (UN Water, 2018). This greater ambition is reflected in both the indicators being measured, going beyond 'improved' to 'safely managed' services (Table 1.1), and the targets, which in most cases require universality in coverage by 2030.[5] The SDGs also incorporated targets for handwashing for the first time, defined as fixed or mobile handwashing facilities with soap and water (Table 1.1). This greater ambition may be

[3] A second issue with the water target, noted by Dar and Khan (2011), occurs where drinking water contaminated by chemicals may cause non-infectious diseases like arsenicosis or fluorosis.
[4] See http://www.un.org/sustainabledevelopment/sustainable-development-goals/.
[5] Unlike other targets which specify 2030, the target for ODF was originally specified for 2025 (Hutton and Varughese, 2016).

necessary to achieve the population health and nutrition improvements long claimed by WASH researchers (Cumming et al., 2019).

The SDGs also reflect an important shift in policy discourse. In addition to including targets for access to basic services, the necessary condition to improve quality of life outcomes, they include use of improved drinking water and sanitation, which is the sufficient condition to improve them. WASH interventions can be conceptualised as containing four components: the technology that is provided to users (e.g., a child's potty and knowledge about safe excreta disposal); the promotional intervention used to encourage demand among the target population (e.g., a government subsidy on the potty purchase price and promotional campaign about excreta disposal) or to improve supply (e.g., capacity building for sanitation providers); the social and physical environment where participants use the technology (e.g., the household and yard); and its suitability for particular groups including disadvantaged people (e.g., children, pregnant women, elderly and disabled people) (Chirgwin et al., 2021).

Improving access to safely managed WASH facilities, and ensuring target populations use them, it is necessary to intervene on both the supply-side – that is, with public and private sector providers of WASH hardware (facilities) and software (know-how) – and on the demand-side – primarily, households and individuals consuming WASH services. Prior to the early-2000s, the focus of WASH evaluation research was principally about understanding, and demonstrating, the efficacy of supply-side interventions to provide WASH technology for household and shared use. Over the last decade or more, the policy debate has increasingly focused on questions about the effectiveness of interventions to promote WASH technology uptake and adherence. Different approaches have been used to promote demand-side behaviour change in the context of water and sanitation provision. For example, directive information and education communication (IEC) through social marketing and subsidies have been traditionally popular means of promoting sanitation and hygiene demand. These have been criticised as inadequate to foster demand to levels required for social benefits, in favour of more participatory methods (e.g., Jenkins and Sugden, 2006; Chambers, 2009).

WASH intervention mechanisms can be defined comprehensively and mutually exclusively (Table 1.2). Mechanisms for providing WASH technologies can be categorised into demand- and supply-side interventions.[6] Demand-side intervention mechanisms include: behaviour change communication (BCC), such as health education and psychosocial 'triggering', for example, social marketing and community-led total sanitation (CLTS); subsidies and microloans for consumers; and legal measures proscribing open defaecation, discharge of contaminated water or dumping of waste (e.g., Cairncross, 1992). For example, psychosocial triggering uses psychosocial factors, principally emotions, like disgust or the desire to be a good parent (Biran et al., 2014) or social pressure, rather than reason, to motivate behaviour change among WASH consumers (de Buck et al., 2017). It aims to promote demand for WASH technology among consumers and may use directive or participatory methods. An example of a directive approach is social marketing, which motivates social change through a combination of product (technology used to meet a need), promotion (to increase desirability and acceptability), place (installation in an appropriate place for users) and price (the cost for users considers affordability) (Cairncross, 2004; Evans et al., 2014). These are often implemented at community level such as in schools and health facilities via approaches such as community health clubs to promote demand (Waterkeyn and Cairncross, 2005). Participatory, bottom-up approaches are also being rapidly scaled up, including participatory hygiene and sanitation transformation (PHAST) in hygiene and community-led total sanitation (CLTS). In CLTS the community is facilitated to discuss how they would like sanitation practices to change, identify problem areas (e.g., 'walks of shame'), and use social cohesion and pressure to motivate people to construct latrines and stop practising open defecation (Kar and Chambers, 2008).

Supply-side intervention mechanisms include: direct provision of technology by an external body (e.g., government, NGO); improving operator performance (e.g., institutional reform, capacity building, operator financing, regulation, and accountability); privatisation (e.g., Galiani et al., 2005) and nationalisation of service delivery; and promoting small-scale independent provider (SSIP) involvement (e.g., sanitation marketing through microloans

[6] I am grateful to Wolf-Peter Schmidt who suggested more clearly differentiating supply- and demand-side interventions.

and capacity building for providers). Direct provision of hardware by an external agency (e.g., government, NGO), covering all interventions where WASH technology (such as a water connection, latrine, water purifier or handwashing facility) is provided at zero capital cost to users (e.g., Feachem et al., 1978). Hardware may be for use in private (household and yard) or public spaces (shared facilities, WASH in health facilities and schools, places of work, commerce, reaction, streets and fields). Measures to improve service provider performance, such as enacting and implementing water quality standards (Cairncross et al., 1996), government regulation of private utility providers (e.g., Ministry of Foreign Affairs, 2011), and reforms to operator financing (e.g., output-based aid or payment-by-results) (Trémolet and Evans, 2010). Encouraging SSIPs like non-profits and the private sector (Sansom et al., 2003) may include microloans for WASH service providers and capacity building. As an example of the latter, sanitation marketing aims to increase availability of sanitation technology and maintenance services (such as pit emptying), by training local artisans to produce sanitation products that are suitable for the varying needs of consumers (e.g., Cameron et al., 2013).

Decentralisation, where community representatives are placed in planning, design, implementation, and operation of the WASH service provider, is an example of an intervention mechanism that combines supply and demand (Poulos et al., 2006). For example, community-driven development (CDD) uses a participatory approach, block grants with cost sharing, and often a component of local institutional strengthening (White et al., 2018). Another approach is the water user association, where management is devolved to the community group while government retains some powers (e.g., Barde, 2017).

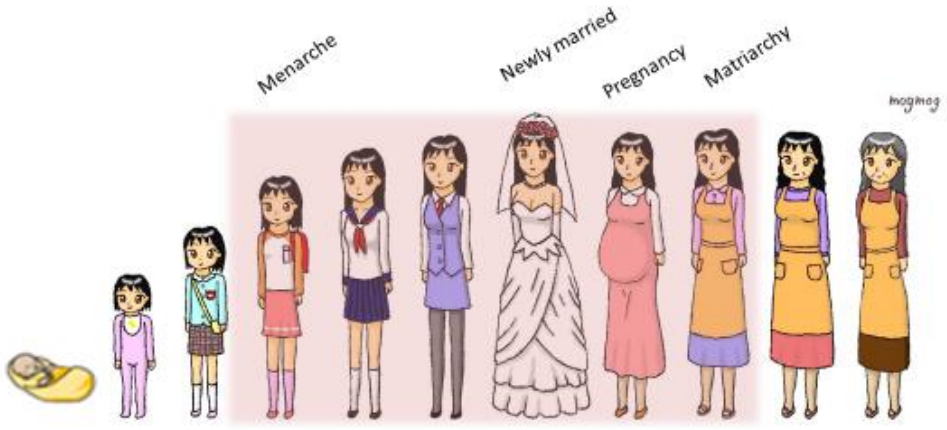
Table 1.2 WASH intervention mechanisms

Intervention type	*Mechanism of delivery*	*Definition*
Demand-side	Health education	Directive hygiene, and sometimes sanitation, education where participants are provided with new knowledge or skills to improve their health based on reasoning. These information campaigns may be provided through television, radio, theatre or printed media; provided directly to specific households or through sessions at community meetings, schools or other places; or provided directly to community leaders or health workers.
	Directive triggering (e.g., social marketing)	Psychosocial 'triggering' covers approaches that use emotional and social cues, pressure, or motivation to encourage community members to change behaviours. Directive mechanisms are typically social marketing campaigns, which use commercial marketing techniques to promote the adoption of beneficial behaviours. They can also include other styles of campaign that use emotional or social triggers rather than information.
	Participatory triggering (e.g., CLTS)	Participatory mechanisms are typically a community-based approach and promote behaviour change through consultation with the community, a two-way dialogue, and joint decision-making. For example, community-led total sanitation (CLTS) uses this mechanism.
	Subsidies and microfinance	All intervention mechanisms that use pricing reform or financial mechanisms to promote the uptake of WASH technologies. This includes subsidies, vouchers, microcredit, and other forms of microfinance, aimed at consumers.
	Legal reform	Intervention mechanisms that enact or implement legal reforms proscribing open defaecation, discharge of contaminated water or dumping of waste.
Supply-side	Direct hardware provision	The provision of any WASH hardware for free and which has been chosen by an external authority. This includes interventions where new or improved water supplies are constructed, handwashing stations are built, soap is handed out, water purifiers given away, latrines provided, or sewer connections installed by external actors (e.g., government or an NGO).
	Improving operator performance	Intervention mechanisms aiming to improve the functioning of the current service provider. This includes improving accountability, oversight or regulation, capacity building and output-based aid.
	Utility ownership	Interventions to change ownership (e.g., privatisation or nationalisation of utilities, public-private partnerships)
	Small-scale independent provider involvement	Intervention mechanisms to encourage small-scale independent organisations, including non-profits, to become the providers of WASH facilities and services on a commercial basis (e.g., sanitation marketing).
Combined interventions	Decentralisation	Focuses on putting the community at the centre of the planning, design, implementation, and operations of their service provider. Examples include community driven development (CDD), also called Social Funds, which are supposed to use a participatory approach to community decision-making, provide block grants with cost sharing, and a component of local institutional strengthening to fully decentralise provision. Other approaches to involving the community but keeping government ownership include water user associations (WUAs).
	Combinations of intervention mechanisms	Intervention mechanisms combining multiple demand-side (e.g., health education with subsidies), supply-side (e.g., hardware provision with privatisation) or combining demand- and supply-side mechanisms (e.g., CLTS and sanitation marketing).

The third important dimension is the social and physical environment where participants interact with the technology. Cairncross et al. (1996) distinguished private domain (dwelling and yard) and public domains (community, schools, places of work, commerce and recreation, fields in rural areas and streets in cities) in disease transmission. The importance of the differentiation is in the potential for communicable disease transmission – the greater potential for single cases to cause epidemics in public spaces – and the types of interventions that are needed to combat transmission – the greater focus on infrastructure investment and regulation in public space, and personal hygiene in private spaces (which also depends on infrastructure investment especially water supply).

The fourth dimension relates to the suitability of WASH technology to different users. For example, women's needs change over their life-cycle, hence WASH service provision needs to be suitable for different points in the reproductive life-cycle, including menarche (e.g., separate toilets for girls at school, promotion of menstrual hygiene management approaches) and maternity (e.g., WASH in health facilities, promotion of hygienic weaning practices) (Figure 1.1).

Figure 1.1 Female reproductive health over the life course



Source: Water Supply and Sanitation Collaborative Council.

Caruso et al. (2017) defined sanitation insecurity as "[i]nsufficient and uncertain access to socio-cultural and social environments that respect and respond to the sanitation needs of individuals, and to adequate physical spaces and resources for independently, comfortably, safely, hygienically, and privately urinating, defecating, and managing menses with dignity at any

time of day or year as needs arise" (p.9). Other disadvantaged or vulnerable groups may also have particular needs, such as water and sanitation facilities for the elderly and infirm, or drinking water treatment for immunocompromised people (e.g., those living with human immunodeficiency virus, HIV). For example, walkways may need to be constructed to prevent falling and elevated seats or rails installed to help elderly people, disabled and pregnant women (ibid., 2017).

1.3 The consequences of limited access to and use of WASH

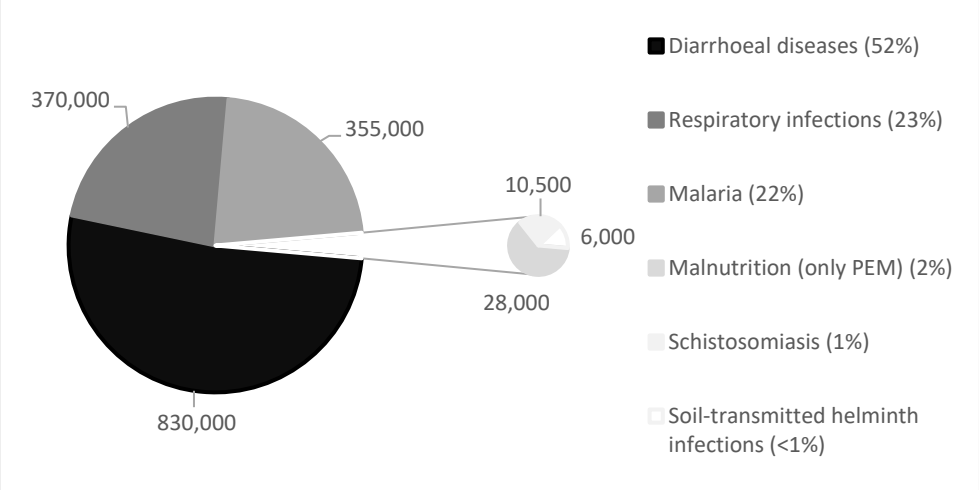
> "Any endeavour to acquire accurate information concerning social influences which may operate prejudicially to health in an area is inseparable from a study of poverty."
>
> M'Gonigle and Kirby (1937, p.22)

Limited, or no, access to safe facilities for eliminating human waste, access to sufficient drinking water, or hygienic washing and food preparation practices exposes individuals to higher levels of infectious disease. Inadequate WASH can contribute to the outbreak and chronic presence of preventable infections like acute lower respiratory tract infections (ARIs) (Rabie and Curtis, 2006) and diarrhoeal disease (Liu et al., 2012; Prüss-Ustün et al., 2019), which are the two biggest killers of children globally (Liu et al., 2012).[7] Enteric disease may also cause tropical enteropathy, a sub-clinical disorder where the lining of the gut wall is damaged by repeated bouts of infection until it is unable to absorb nutrients adequately (Shiffman et al., 1978; Humphreys, 2009). Chronic high enteric infection rates are among the leading causes of undernutrition and death in children in developing countries (Cairncross et al., 2014). According to recent Global Burden of Disease estimates (Prüss-Ustün et al., 2019), inadequate WASH is associated with 1.6 million deaths per year, due to diarrhoea, acute respiratory infection, malnutrition due to protein energy management (PEM) and, because of water mismanagement, malaria (Figure 1.2).

[7] Hygiene and water supply are also likely to be key blocks to the transmission of COVID-19, a type of acute lower respiratory tract infection (Howard et al., 2020).

Diarrhoea alone kills 850,000 people every year, 300,000 of whom are children aged under 5 (Prüss-Ustün et al., 2019). Each death is a personal tragedy (White, 2004). Parasitic worm infections, associated with inadequate sanitation (e.g., schistosomiasis), are responsible for 39 million disability-adjusted life years (DALYs), equivalent to the global burden of mortality for malaria and tuberculosis combined (Stephenson et al., 2000). Trachoma, a water-washed eye infection causing blindness, spread by the *Musca sorbens* fly which breeds in human excrement, affects an estimated 146 million people worldwide (Ejere et al., 2012). Water supply changes may also affect rates of arsenic poisoning due to groundwater consumption, which can cause nutritional deficiency, cancer and death (Dar and Khan, 2011; Jones-Hughes, 2013).

Figure 1.2 Estimated annual global deaths due to inadequate WASH



Source: data from Prüss-Ustün et al. (2019).

There may also be important externalities from private consumption of improved WASH services through environmental health spillovers (Root, 2001; Barreto et al., 2007; Spears, 2013; Duflo et al., 2015), operating in private (household and yard) and public (places of work, education, commerce, recreation, street and fields) domains (Cairncross et al., 1996). For example, the World Bank (2008) estimated environmental costs of poor sanitation at 2 per cent of GDP in South Asia (Cambodia, Indonesia, the Philippines and Vietnam). In sum, water-related diseases are responsible for an estimated 21 per cent of the global disease burden (Black et al., 2010). Poor access in places with high population density, may explain why some

countries, particularly in South Asia, have worse child malnutrition outcomes than their income levels alone would predict (Spears, 2013).

Beyond the potentially life-threatening consequences of ARIs and enteric infections, poor access and use of WASH may also affect social and economic outcomes, both directly and through follow-on effects. This may include diminished educational attainment (Hennegan and Montgomery, 2016). For example, a multi-country study in sub-Saharan Africa found that millions of children were tasked with collecting water (especially girls) for journey times greater than 30 minutes (Graham et al., 2016), likely affecting their education. Where female adults are required to collect the water, which is most cases, older children may be pulled out of the school to care for younger ones (Koolwal and van de Walle, 2010). Diminished educational attainment, due to children's school enrolment and attendance as well as teacher attendance, as well as delayed entry to the labour market, have implications for employment, life-time wage earnings and income (Poulos et al., 2006; Hutton et al., 2007).

While all suffer loss of dignity from open defecation and drudgery from water collection, women and girls suffer particularly. Women do most of the water carrying when households lack access to an improved water source in Africa and Asia (Sorenson et al., 2011). Originally, McSweeney (1979) had reported that the burden of time spent on domestic chores in Burkina Faso started in a girl's childhood, was around 7-8 hours per day by age 9 (double that of boys of similar age) and women and girls were responsible for all the water collection. Feachem et al. (1978) estimated that 96 percent of water collections in Lesotho were made by women and girls. Cairncross and Cliff (1987) reported time savings associated with water supply improvements for women in Mozambique, which were put to other household activities (food preparation and childcare), suggesting a possible mechanism through which WASH impacts on nutrition. Women and girls still did most of the water collection in analysis of DHS for 24 sub-Saharan Africa countries by Graham et al. (2016). Other important consequences include musculoskeletal injuries from repeated heavy load carrying (Porter et al., 2013). For example, women interviewed after water supply improvement in a slum in Gujarat, India, said that not having to carry buckets of water, "apart from saving time and labour, has reduced their back problems" (Parikh and McRobie, 2009, p. 276).

People risk becoming road casualties, and risk attack and assault by 'pests and perverts' (Campbell et al., 2015). For example, Cairncross and Cliff (1987) found in northern Mozambique that, when the functioning village standpipe broke down, women were forced to rely on traditional sources. The choice included a water source 8 km away, taking between 4 and 7 hours (travel time and queueing) for the return journey, or one 4 km away, where "[a] few women spent the night… despite the danger of lions, waiting for water to appear in the holes dug for that purpose" (p.51). Control over water supply and who does the collecting for household use remains highly gendered. As noted by (Thompson et al., 2001, p.63): "[i]t may be a male decision to install piped water to a village, but the women often have to operate and maintain the water supply and deal with problems when it fails. In fact, in many places, it would seem shameful for a man to be seen collecting the family's water supply."

Women and girls may face danger when they have to wait until after dark to urinate or defecate with privacy (Sorenson et al., 2011; Sommer et al., 2014; Sahoo et al., 2015; WaterAid, undated). For example, studies in Kenya (Winter and Barchi, 2016) and India (Jadhav et al., 2016) found that women who openly defaecated were more likely to experience non-partner sexual and/or physical violence; and in India, twice as many women who openly defaecated experienced non-partner violence than those with a private toilet. They also experience hardships where inadequate WASH facilities constrain menstrual hygiene management causing urinary tract infections (Torondel et al., 2018) and absence from school and work (Sumpter and Torondel, 2013). There may also be adverse maternal and child health implications due to inadequate WASH services in health facilities and other places of new-born delivery (Benova et al., 2014). Pregnant women and neonates are thought to be a particularly high-risk group because infection and sepsis are major causes of maternal and neonatal mortality (Liu et al., 2012). Campbell et al. (2015) systematically mapped a range of possible consequences for maternal health due to contact with contaminated water (e.g., arsenicosis, schistosomiasis, hepatitis E), and availability of water (e.g., malaria, uterine prolapse due to water carrying), sanitation (e.g., rape), and hygiene (e.g., influenza). More generally, disadvantaged groups, such as women, children, the elderly and people with disabilities, are less likely to have access to appropriate WASH technologies (whether drinking water supplies of

sufficient quantity and quality, means of safe excreta disposal, and hygiene practices), and therefore more likely to experience negative health and socioeconomic consequences.

Other longer-term economic implications arise due to delayed entry to the labour market, and monetary losses due to costs of medical treatment and aversion costs of treating and storing unclean water or purchasing water from vendors (Cairncross and Kinnear, 1992; Bosch et al., 2002). These costs can be exorbitant for poor households in urban informal settlements (slums) who are unserved by house connections. For example, the costs of vendor supply were estimated at 7-11 times higher than public utility water supply in Nairobi, Kenya, 12-25 times in Dhaka, Bangladesh, 28-83 times higher in Karachi, Pakistan, 17-100 times higher in Port-au-Prince, Haiti, and 100 times higher in Nouakchott, Mauritania (Bhatia and Falkenmark, 1993, p.14). In a study in Khartoum, Sudan, where up to 56 percent of household income in squatter areas was spent on vendor water (Cairncross and Kinnear, 1992), the income and price elasticities of demand for water were found to be very inelastic (that is, demand is relatively unresponsive to changes in income and price). It was therefore suspected that the poorest households would need to substitute food expenditure to meet water needs, causing malnutrition.

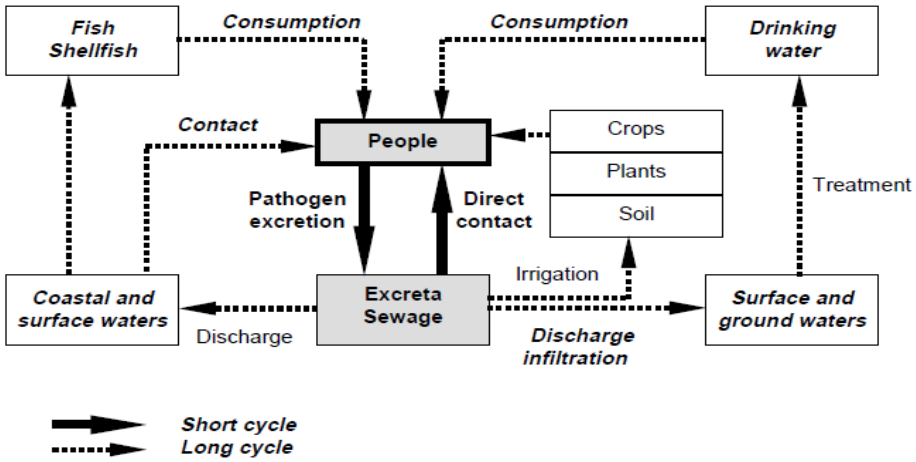
For all these reasons, improving WASH service access and use is likely to support conditions for virtuous cycles of development and pro-poor growth (Ramirez et al., 1998; Anderson and Waddington, 2007). What remains at issue, however, is the extent of evidence supporting these claims and the magnitudes of the possible impacts of WASH interventions in particular contexts and for groups of participants.

1.4 Linking WASH technology interventions and outcomes

> "There is present here an important field of research which has been left almost unexplored. How far and in what respects are the common defects of childhood associated with the disabilities of adult life?"
>
> M'Gonigle and Kirby (1937, p.52)

As noted in Cairncross et al. (1996) and later Bosch et al. (2002), water-related disease transmission operates through two main routes: direct transmission through the private domain or 'short cycle' due to poor personal hygiene; and indirect transmission through the public domain or 'long cycle' due to environmental pollution (Figure 1.3).

Figure 1.3 Pathways of human exposure to water-related pathogens



Source: Bosch et al. (2002).

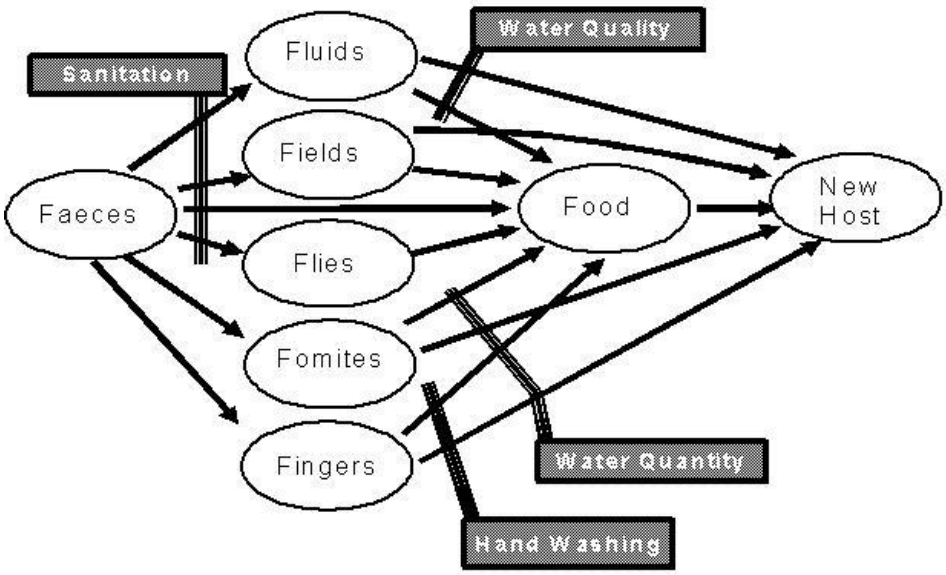
Breaking the long cycle requires community infrastructure investment, such as lined latrine pits to prevent contamination of ground water, and sewage treatment to prevent contamination of coastal and surface water (e.g., transmission between humans and shellfish of gastroenteric infections like norovirus[8]). Breaking the short cycle requires changes in personal behaviour and practices mainly in the household.

Figure 1.4 shows a theoretical depiction of the direct communication of faeco-oral pathogens between individuals (Wagner and Lanoix, 1958). Later called the 'F-diagram' (e.g., Kawata, 1978), it shows the behavioural transmission routes for various water-related diseases from faeces to future hosts via water (fluids), hands (fingers), arthropods (flies), soil (fields) and food. A sixth transmission route has since been identified, 'fomites' – that is, objects acting as disease-carrying vectors such as clothes, utensils, toys and furniture (Cairncross and Feachem, 2018). Implicit in the figure are three

[8] The Guardian, January 6, 2020: Brittany oyster farms hit by gastroenteritis epidemic. https://www.theguardian.com/world/2020/jan/06/brittany-oyster-farms-gastroenteritis-epidemic-sewage (accessed 6 January 2020).

water-related, faecal-borne disease transmission routes: water-borne diseases transmitted through ingesting infected water and water-washed diseases transmitted through inadequate drinking water supply and hygiene (e.g., cholera, diarrhoeal disease, hepatitis, typhoid), and water-based diseases transmitted by penetrating skin (e.g., schistosomiasis transmitted in water, and *Ascaris*, hookworm and whipworm in contaminated soil). The F-diagram focuses on faecal-borne diseases, but additional water-related infections that are not faeces-related exist through the water-washed route such as respiratory infections, especially through hand hygiene (Rabie and Curtis, 2006), fomites (e.g., Levy et al., 2013), skin and eye infections (e.g., trachoma, scabies) and louse-borne infections (e.g., typhus), and water-based diseases that are transmitted through ingestion (e.g., Guinea worm disease) (White et al., 1972, p.163). A fourth transmission route is water-related insect vectors, which pass on disease by biting near water (e.g., sleeping sickness) or breeding in water (e.g., chikungunya, dengue, malaria).

Figure 1.4 The 'F'-diagram showing faecal-oral disease transmission



Source: Cairncross and Feachem (2018).

Figure 1.4 shows sanitation as a primary barrier to faecal-related disease transmission, when excreta carrying faecal pathogens are eliminated from the environment or human consumption. Primary barriers also include handwashing and water quantity, important for stopping transmission primarily in the domestic domain (fingers and fomites). Due to faecal

contamination of drinking water between source and point-of-use (POU), hygienic approaches may be needed to store clean water collected at source or treat water for contaminants in the household at POU (Wright et al., 2004; Fewtrell and Colford, 2004). Better access to water supply (quantity) may improve health by reducing contamination in the environment by enabling better personal hygiene (e.g., handwashing) and environmental hygiene (e.g., safe disposal of faeces). The secondary barrier is drinking water quality (Kawata, 1978). Factors such as environmental faecal contamination may prevent impacts from clean drinking water provision being realised due to the amount of time infants and children, who are the most susceptible to diarrhoeal disease, spend on the floor and putting their fingers in their mouths (e.g., Cattaneo et al., 2009).[9]

Outcomes of WASH sector interventions can be categorised into six main groups: intermediate outcomes relating to WASH access, knowledge, attitudes and behaviours (e.g., time use, consumer satisfaction, environmental pathogen contamination); health outcomes due to water-related health infection (e.g., diarrhoeal morbidity, acute respiratory infections, gastro-intestinal worm infections); other health outcomes, which are largely gendered (musculoskeletal disorders, reproductive tract infections, injuries and psychosocial health); nutritional status, relating to water-related disease and carer and children's time use; mortality, particularly in childhood; and socioeconomic outcomes (e.g., education and cognitive development, net earnings).

A conceptual framework linking WASH interventions with outcomes along the causal pathway is depicted in Figure 1.5. The framework was developed based on a review of the academic and policy literature, and in consultation with researchers, WASH practitioners and WASH programming organisations (Chirgwin et al., 2021). Intervention mechanisms are presented to the left of the figure: on the supply side, water and sanitation hardware provision by external agencies, improved operator performance, private sector participation and contracting out, and decentralisation; and on the demand side, behaviour change communication, pricing reforms and

[9] The F-diagram relates to faecal-borne pathogen related disease transmission. Non-infectious waterborne diseases, such as arsenicosis and fluorosis, caused through chemical contamination of water, are increasingly recognised as a source of human morbidity and mortality (Dar and Khan, 2011).

financial support. Quality of life outcomes – water-related health, other health and socioeconomic impacts – are presented on the right. Outputs are defined as the direct consequences of WASH provision and outcomes as depending on participant behaviour. Outputs providing access to WASH are mainly technological, whereas outcomes are behavioural. However, some intervention mechanisms aim to stimulate access by encouraging behaviour (e.g., construction of latrines or wells), so the distinction is not always clear cut. The causal pathway, therefore, shows the stages that interventions are turned into impacts (quality of life outcomes), through activities (construction of new facilities, behaviour change campaigns), outputs (better access to, quality of, knowledge of, and attitudes towards WASH services and practices) and intermediate outcomes (behaviour change relating to access and use of improved WASH services).

Figure 1.5 is highly simplified and excludes underlying assumptions. Links in the causal pathway between interventions and outcomes are not automatic. For example, water treatments may not lead to less faecal contamination if the treatment technology itself is not efficacious in combating parasitic infections (Arnold and Colford, 2007). An example would be chlorination which is not effective against cryptosporidium, a common cause of diarrhoeal morbidity and mortality, especially among immunocompromised groups such as those living with HIV (Havelaar et al., 2003, cited in Abubakar et al., 2007). And even an efficacious technology may not reduce contamination if used improperly, for example where insufficient protective agents are applied to treat drinking water, or insufficient time available to purify water before ingestion. In the case of drinking water provided at source, there may be environmental contamination during transport (e.g., use of contaminated storage containers) or poor personal hygiene at point-of-use (e.g., when contaminated hands are put in water storage containers) (Wright et al., 2004). Other factors limiting effectiveness are due to adoption, for example users may dislike the odour and taste of chlorinated water.

Similarly, providing latrines may not necessarily lead to less open defaecation, for various reasons such as the quality of facilities (cleanliness and smell) or concerns from pit owners about the frequency that the pit will need to be emptied. Nor may latrine provision lead to better health and nutrition if open defecation is still practised by some people in densely

populated areas (Kar and Chambers, 2008). Latrines are not usually designed for or used by children, who may be afraid of going into dark places or of falling into the pit. This may be particularly problematic for reducing environmental contamination because children's excreta are more likely to contain infectious pathogens than adults' (Majorin et al., 2019), even though they may not be thought dangerous or offensive (Curtis et al., 1995).

Preventive technologies tend to be adopted more slowly as benefits are difficult to observe (Rogers, 2005). This applies particularly to WASH technologies whose main benefit is to reduce diseases, the prevalence of which may typically be infrequent (or effects unobserved) outside of epidemics. For example, the incidence of diarrhoeal disease among study participants in L&MICs was around 10 percent in one systematic review (Waddington et al., 2009). An average reduction in risk of child diarrhoea by 30 percent, the typical pooled effect size found in meta-analyses of WASH technology evaluations, would therefore only reduce the number of diarrhoeal days from 10 to 7 percent on average, if the measure were based on prevalence.[10] Even a reduction in average risk by 50 percent for household water filtration, would reduce the typical child diarrhoeal risk from 3 episodes per year to 1.5 episodes (Clasen et al., 2015). In contrast, where the benefits of a technology are easily observed by those directly affected, such as poor women and children collecting water every day, and hence adoption likely to be rapid where it can be adequately provided, it is more likely that underinvestment in the technology would be explained by systemic undervaluation of the benefits and costs (including opportunity costs) for the affected groups, both by public authorities and household decision-makers. Indeed, as discussed later in this Thesis (Chapter 3, Section 3.5.2), while health is the main preventive outcome for WASH, it is not a major motivating factor for WASH behaviours.

[10] Diarrhoeal illness is usually measured as the risk, incidence or prevalence. Risk measures the probability of being ill during the measurement period. Incidence density or rate measures the average risk over the measurement period measured in average number of discrete disease spells, where a spell is usually demarcated by at least three intervening diarrhoea-free days (Bacqui et al., 1991). Longitudinal prevalence is more closely associated with duration of illness, usually measured as the proportion of days of illness during the measurement period. Longitudinal prevalence of illness is preferred on theoretical grounds and empirically is more strongly associated with child mortality and weight gain than incidence (Morris et al., 1996). Different technologies may also affect measures of incidence and prevalence differently. For example, hygienic practices such as removal of faeces from the yard may have greater impact on spell duration (Gross et al., 1989).

Figure 1.5 WASH interventions simplified causal pathway

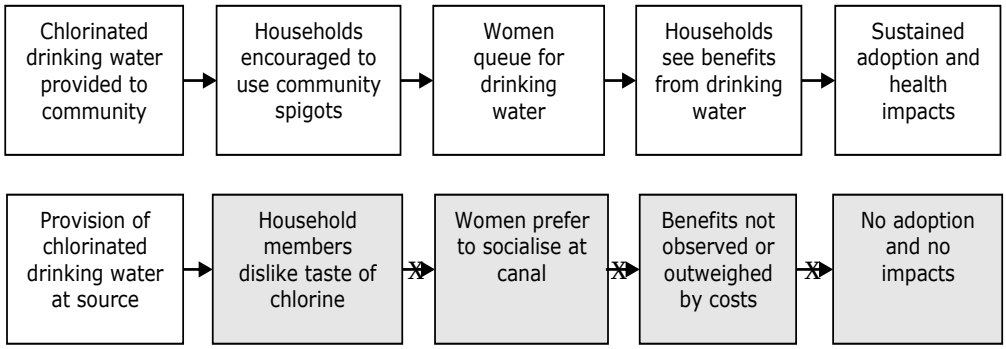
Interventions	Activities	Outputs	Outcomes	Impacts
Demand-side ○ Behaviour change communication (e.g., PHAST, CLTS, social marketing) ○ Pricing reforms (tariffs and subsidies) and financial support (e.g., microcredit) ○ Legal reform (e.g., against open defaecation or dumping of waste) **Supply-side** ○ Direct hardware provision by an external agency (e.g., government, NGO) ○ Privatisation or nationalisation of service delivery ○ Small-scale independent provider involvement (e.g., sanitation marketing) ○ Improving operator performance (e.g., regulation) **Demand- and supply-side** ○ Decentralisation (e.g., community-driven development, water user associations) ○ Combined demand and supply interventions (e.g., CLTS with sanitation marketing) **WASH technology; place of use** Water, sanitation and hygiene technology for use in households, schools or health facilities, in rural, urban, informal (slum) communities and refugee camps **Target group** WASH over the life-course and vulnerable groups.	○ Water supply facilities construction and maintenance ○ Water treatment provision and maintenance ○ BCC about drinking water ○ Hygiene facility construction and maintenance ○ BCC about hand, food and personal hygiene ○ Sanitation facility construction and maintenance ○ BCC about sanitation	○ Better, more reliable access to drinking water supply ○ Better quality drinking water ○ Drinking water knowledge and attitudes ○ Better access to hygiene facilities ○ Hygiene knowledge and attitudes ○ Access to safe sanitation ○ Sanitation knowledge and attitudes	○ Increased use of water ○ Less contamination (water-washed infection) and exposure to insect vectors ○ Time-savings ○ Consumer surplus ○ Less contamination of drinking water (water-borne infection) ○ Improved hygiene practices (hand and food hygiene, including infant weaning) ○ Less contamination (water-washed faeco-oral, respiratory, skin, eye and louse-borne) ○ Use of sanitation facilities ○ Reduced open defecation ○ Less contamination (water-washed, water-borne, water-based)	**Water-related ill-health** ○ Less water-borne faecal-oral (e.g., diarrhoea) and chemical-related disease (e.g., arsenicosis) ○ Less water-washed and water-based faecal-oral (e.g., diarrhoea), water-related disease (e.g., ARIs) and insect vector disease (e.g., malaria) **Other health** ○ Reduced musculoskeletal disorder ○ Fewer injuries ○ Less reproductive tract infection ○ Improved psycho-social health (safety, stress, dignity, happiness) **Improved nutrition** ○ Improved child growth ○ Less anaemia ○ Less enteropathy **Improved survival** ○ Fewer infant and child deaths **Socio-economic benefits** ○ School enrolment, attendance and attainment ○ Higher income and consumption ○ Lower health care and aversion costs **Inequality in impacts** Impacts for disadvantaged and marginalised groups (e.g., children, poor people, pregnant women, people living with disability and HIV)

Source: Chirgwin et al. (2021).

Sustaining impacts and achieving them at scale requires the continued wide acceptance and adoption of new technology, which may require additional promotional approaches. Sustainability and scalability of impacts are therefore central issues for policy and practice. Sustainability of impacts requires continued adherence by beneficiaries, solutions to 'slippages' in behaviour and financial barriers to uptake, as well as technical solutions to ensure service delivery reliability. Scalability requires that impacts measured in small-scale efficacy settings (the 'ideal settings' measured in many field trials) are achievable in the context of programme effectiveness ('real world' settings) where fidelity of implementation becomes crucial (Bamberger et al., 2010). For example, hygiene information, education and behaviour change activities are usually a component of most, if not all, programme designs which aim to scale-up service provision. However, there are concerns about whether these activities are being implemented in practice (Jimenez et al., 2014).

However, the effectiveness of WASH technology in preventing disease transmission depends on both the biological efficacy of the technology and its acceptability and use, or effectiveness, among consumers in the environment where it is based (Eisenstein et al., 2007). Acceptability and use in turn are determined by the WASH intervention mechanism, which motivates behaviour change by triggering drives (e.g., disgust), emotions (e.g., status) or interests (e.g., curiosity) (Aunger and Curtis, 2016). Authors of diarrhoea efficacy studies have referred to lack of convenience and limited observability of health benefits in explaining why compliance rates may be low for household water treatment (Quick et al., 2002). Rogers (2005) documented the low level of use of public spigots in 1960s Egypt, despite government media campaigns warning people of the risks from drinking canal water. Qualitative research suggested various causes, including that users did not like the chemical taste of the chlorinated water, rumours that the chemicals were being used to control fertility, women preferring to gather water from the canal banks where they socialised, and long queues, and fighting in the queues, due to low water pressure (Figure 1.6).

Figure 1.6 Programme theory and practice – public spigots in Egypt



Source: author drawing on the description contained in Rogers (2005).

1.5 WASH sector impact evaluation

> "It was decided that the original area was too large to be dealt with under one scheme, and it was therefore divided into two portions. For convenience a line of division was decided upon which ran along a street called 'Smithfield'... It will be seen that the conditions were very favourable for investigation. There was in the first place, a population transferred from slum dwellings to a modern, self-contained housing estate, and kept intact without admixture with other populations. There was, further, a second population that continued to dwell in slum houses and served as a control."
>
> M'Gonigle and Kirby (1937, pp.108-9)

Impact evaluations quantify the net effect of providing an intervention to a group on measured outcomes, with reference to a counterfactual group that receives no, or a different, intervention (Cairncross et al., 1980; Briscoe et al., 1986; Shadish et al., 2002; Duflo et al., 2006). 'Rigour' in impact evaluation is usually defined in relation to the ability of the study to measure the relationship in an unconfounded way. There is a long history of programme evaluation in WASH, and the types of studies thought suitable, and even possible, in WASH evaluation has changed over time. In the 1970s, a World Bank expert panel had stated that "long-term longitudinal studies of large size and expense are probably the only means through which there is any chance of isolating a specific quantitative relationship between water supply and health" (World Bank 1976; quoted in Churchill et al., 1987). Randomised

controlled trials were thought to be overly costly and time-consuming and the panel recommended the World Bank not to fund such studies.

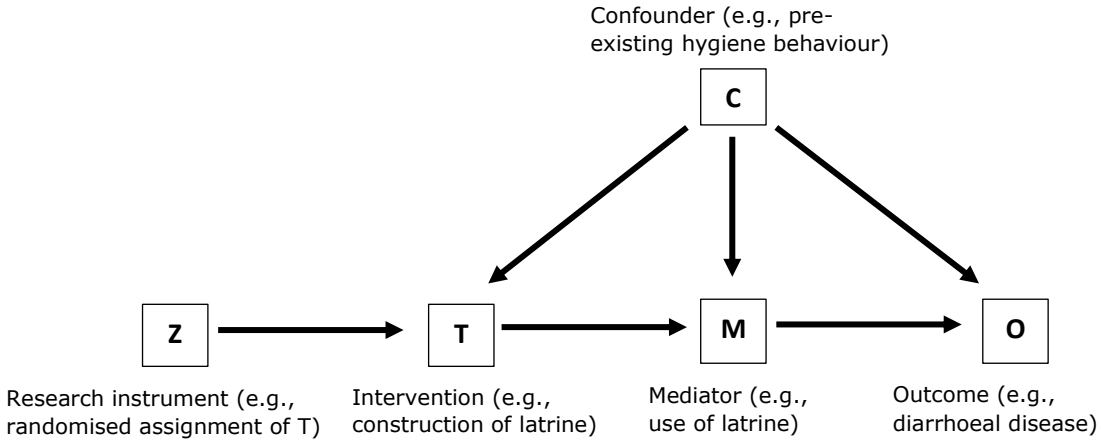
In 1970, a study compared diarrhoeal disease before and after hygiene education was provided and bore-hole pits dug to dispose of children's faeces, in villages in India (Kumar et al., 1970). However, the earliest controlled impact evaluations of WASH interventions in L&MICs appear to be Feachem et al.'s (1978) study of the effects of rural water supply provision on behaviour, income and health in Lesotho, and a study of piped water and hygiene promotion by Shiffman et al. (1978) in Guatemala. Torún (1982) also conducted an early clustered evaluation of piped water supply provision and health education in two villages in Guatemala, one of which had received the intervention. Prospectively designed factorial trials conducted in field settings with contemporaneous measurement from pre-test in at least two groups that receive different interventions, have been published in L&MICs since Khan (1982) followed up individual shigellosis cases in households in Bangladesh. Khan (1982) divided participants into four groups – three that were provided soap, handwashing pitchers, or both soap and handwashing pitchers, and a control group that received no soap or pitchers – to investigate measures to prevent disease transmission in the private domain.

Standards for evaluation in water supply and sanitation were articulated early on. Briscoe et al. (1985, 1986) helped inform the 'first generation' of WASH health impact evaluations, by articulating methods to quantify the effects of WASH service provision, usually on diarrhoeal disease, using randomised and non-randomised approaches. It is usually thought necessary to collect study participant data contemporaneously against a control group (called comparison group in non-randomised studies) to control for confounding – that is, changes in outcomes caused by factors other than the treatment. For example, Figure 1.7 presents the causal pathway from the exposure or treatment (T), access to latrine, and mediator (M), use of latrine, through to outcome (O), diarrhoea, using a directive acyclic graph (DAG) (e.g., Hernán et al., 2004; Pearl and McKenzie, 2018). It also shows one of the potential confounders (C) in the relationship, pre-existing hygiene behaviour. In theory, the unbiased causal relationship between exposure and outcome can be estimated in multivariate (or stratified) analysis by controlling on (stratifying by) pre-existing hygiene behaviour and any other factors that may simultaneously determine intervention exposure and

diarrhoea such as socioeconomic status and water supply access and use (Cairncross and Kolsky, 1997), provided these can be measured reliably.

Unfortunately, factors determining whether individuals and groups participate in, or benefit from, interventions are often innate and unobservable. For example, intervention sites may be chosen by planners for political reasons, because they are accessible, or perhaps because they are the neediest (programme placement bias). Participation by individuals and households in treatment take-up and adherence is usually voluntary and determined by non-random factors like socioeconomic status, attitudes or individual self-efficacy (self-selection bias). These factors are usually unknown or can only be measured with error. Prospective randomised assignment to intervention, where feasible and ethical, is usually the preferred approach for causal identification (Sacks et al., 1982; Briscoe et al., 1986; Habicht et al., 1999; Shadish et al., 2002; Duflo et al., 2006). Shown as *Z* in Figure 1.7, randomisation by nature is uncorrelated at baseline with confounders that determine exposure (latrine construction), adherence (use of latrine) and changes in outcome (diarrhoea). In contrast, pre-existing hygiene behaviour, which may be impossible to observe without bias (especially in a retrospective study without baseline measurement), is likely to confound the relationships between intervention participation, adherence and disease outcomes (Figure 1.7).

Figure 1.7 Confounding of the causal pathway for latrine access



Confounding becomes more problematic further along the causal pathway (Cairncross and Kolsky, 1997; White, 2014). Therefore, even in well-implemented RCTs, it can be difficult to measure 'endpoint' outcomes like

child linear growth with precision, as was found, for example, in the recent WASH-Benefits trial in Bangladesh (Luby et al., 2018). There may also be confounding due to selection bias. For example, if an intervention is sufficiently protective against ill-health to reduce death, a perverse effect may be estimated on ill-health and nutrition outcomes, if the weakest children are saved in the intervention, children who otherwise die in the control population (Lee et al., 1997). In addition, due to the longer causal pathway – and especially when combined with imperfect take-up and adherence – the effects of WASH promotional interventions may not be detectable with statistical precision for final quality of life outcomes.

Although controlled field trials using more rigorous designs with larger samples have been available since Kirchhoff et al.'s (1985) placebo-blinded crossover trial of water chlorination in rural Brazil, the first RCTs of WASH in L&MICs were not published until Austin's (1993) study of household drinking water treatment by sodium hypochlorite on diarrhoea morbidity in the Gambia, and the *Universidad Rafael Landívar* (URL, 1995) study of household filtration in Guatemala.[11] Since that time, RCTs of water treatment interventions have become more common (Clasen et al., 2015), including double-blinded trials of the impact of household water treatment on carer-reported diarrhoea (e.g., Boisson et al., 2013).

RCTs were only thought practicable for evaluations of small-scale technologies like household water treatment and handwashing with soap, due to the high costs inherent in conducting clustered trials of water supply and sanitation at scale (Cairncross et al., 2014). However, reflecting the policy debate around the effectiveness of interventions to promote WASH technology uptake and adherence, and new resources made available, especially by The Gates Foundation, there has been an associated increase in production of evaluations of WASH intervention mechanisms. This 'second generation' of WASH impact evaluation research focuses on measuring behaviour change and broader health and socioeconomic outcomes, including, for example, large-scale cluster randomised studies of the Indian government's Total Sanitation Campaign (Clasen et al., 2014), community-

[11] Initially the RCTs were almost exclusively for studies of point-of-use water treatment. A famous set were carried out by the Centers for Disease Control and Prevention (CDC) with funding from Procter and Gamble, who make chlorine as well as soap (Sandy Cairncross, pers. comm.).

led total sanitation in Mali (Pickering et al., 2015) and school-based water supply, latrines and handwashing in Kenya (e.g., Freeman et al., 2012).

Briscoe et al. (1986) provided standards for non-randomised methods of impact evaluation. These included 'quasi-experimental designs' – prospective non-randomised studies where the investigator collects data from treatment and comparison groups as part of the study, as well as retrospective case-control designs for rarer outcomes like mortality.

Examples of non-randomised approaches include:

- Studies with assignment of units based on practitioner or participant selection and contemporaneous measurement of outcomes by investigators at pre-test and post-test in treatment and comparison groups,12 or contemporaneous measurement by investigators in treatment and comparison group at post-test only. These include studies that use methods such as statistical matching on baseline characteristics and/or direct control for confounding in adjusted analysis (e.g., Reese et al., 2019). The more rigorous approaches compare communities receiving an intervention to a geographically separate comparison group without access to the intervention, rather than comparing those within eligible communities based on self-selected participation (e.g., Gross et al., 1989).
- Non-randomised crossover trials where treatment and comparison are swapped after a certain time (e.g., Kirchhoff et al., 1985).
- Non-randomised studies (NRS) designed retrospectively – that is, after the intervention has occurred – using cross-section data (e.g., Khan, 1987) and case-control (e.g., Victora et al., 1988).

Non-randomised approaches to causal identification also exist that can control for unobservable confounding, including so-called 'as-if randomised' studies, like natural experiments (Figure 1.8). Like RCTs, as-if randomised designs are based on knowledge about allocation rules that are external to participants. Causal identification in these studies rests on the assumption that the factors determining assignment are not caused by the outcomes of interest nor are correlated at baseline with its other determinants, or it can

[12] This designation also applies to RCTs with non-compliance that are analysed using treatment-on-the-treated analysis (also called average treatment effect on the treated, ATET).

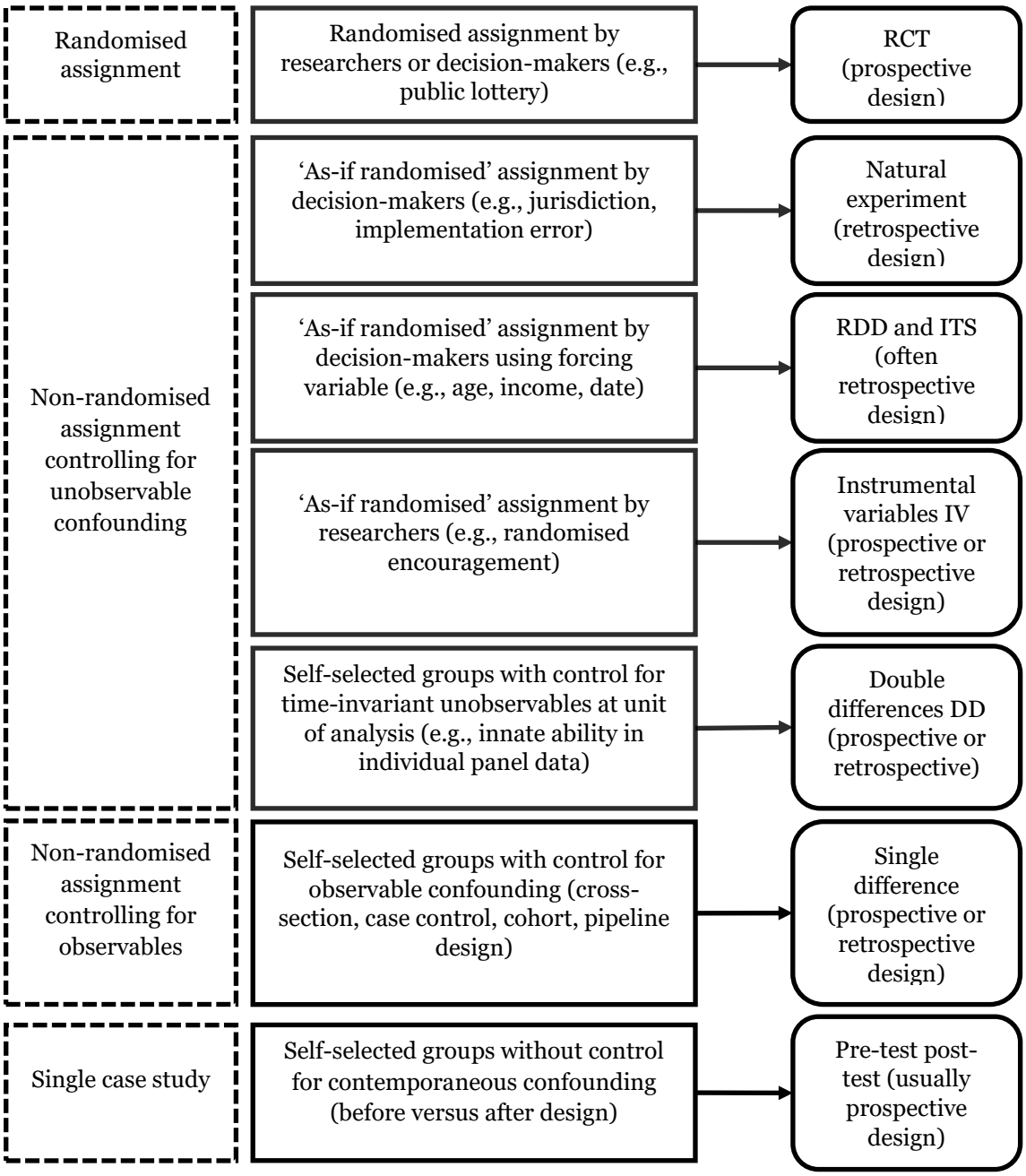
be credibly modelled in analysis. Examples of non-randomised approaches with selection on unobservables:

- natural experiments in which treatment is assigned quasi-randomly by decision-makers using an exogenous mechanism such as an arbitrary allocation of the water supply (e.g., Snow, 1855);
- regression discontinuity design (RDD) in which assignment by decision-makers is based on a threshold on an ordinal or continuous variable (e.g., test score, age or date), where quasi-random variation can be determined close to the treatment threshold (Villar and Waddington, 2019); similarly, interrupted time series (ITS), where repeated measurements are in intervention groups before and after treatment has been allocated (e.g., Barreto et al., 2007). RDD and ITS are often undertaken retrospectively as natural experiments using observational data; for example, village water supplies in Guinea (Ziegelhöfer, 2012) and India (Duflo et al., 2015) and financial incentives for achieving open defaecation free (ODF) villages in India (Spears, 2013);
- instrumental variables (IV) estimation in which quasi-randomly distributed exogenous factors can be identified, often retrospectively, which are correlated with treatment assignment but do not directly determine outcomes. For example, topography has been argued to fulfil these criteria, such as land gradient in studies of the effects of dams on poverty in Kerala, India (Duflo and Pande, 2007) and water treatment plants on diarrhoea and nutrition (Zhang, 2011). IV is also done in prospective evaluations of interventions that would be difficult or impossible to conduct under controlled conditions. For example, where programme eligibility is universal, a pure controlled study design is not possible. However, marketing information about the programme can be randomly assigned (randomised encouragement design), as in the evaluation of a programme providing credit to households for piped water connections in urban Morocco (Devoto et al., 2012); and
- double differences (DD) estimation applied to longitudinal panel data conducted at intervention pre-test and post-test, for example investigation of water supply in peri-urban Argentina using panel data or pseudo-panels of repeated cross-sections with an intervention and comparison group (Galiani et al., 2007).

The first three methods can account for both time-varying and time-invariant sources of unobservable differences between participants and comparisons.

Where allocation rules are not observable, possible sources of confounding must be modelled in statistical analysis. In the special case of DD, control for time-invariant unobserved confounding is possible. Single difference estimation of cross-section or cohort data using multivariate regression or statistical matching to control for measured confounders directly, is not generally able to control for time-varying or time-invariant factors.

Figure 1.8 Study designs to quantify treatment effects



Source: adapted from Waddington et al. (2012, 2017).

Single difference estimation may still provide valid estimates of impact for certain types of outcomes where: 1) methods are used to match groups

statistically on observable factors collected at baseline, which can be credibly argued as strongly correlated with unobservable sources of confounding; or 2) when theory of change analysis of intermediate outcomes and causal mechanisms supports estimates of final outcomes. Examples include:

- comparison group designs employing statistical matching methods (e.g., propensity score matching, PSM), often based on retrospective analysis of household survey data in analysis of household water and sanitation in India (Jalan and Ravallion, 2003; Geruso and Spears, 2018), or case-controls using matched health facility administrative data in investigation of latrines in Lesotho (Daniels et al., 1990a);
- cohort designs that control for observable confounders and estimate impacts on outcomes along the causal pathway (e.g., Ercumen et al., 2015b; Reese et al., 2019); and
- uncontrolled pre-test post-test (before versus after) designs and pipeline designs where changes are measured a short period of time following the intervention, or the causal pathway is short, where the expected effect is large, and confounding is unlikely (Victora et al., 2004). A good example of this is time-savings outcomes (e.g., Cairncross and Cliff, 1987).

There remains a need for rigorous observational approaches to evaluate impacts over the very long term, because it is difficult to prevent control group contamination or locate individuals for follow-up in prospective studies. This includes long-term outcomes potentially taking decades to materialise, like adult earnings potential in response to WASH conditions in childhood, and long-term interventions like establishing "a sanitation market offering good products and to persuade people that a latrine can make their life, cleaner and healthier, or even be a sign of social status" (Schmidt, 2014, p.524). As noted above, there is great policy interest in impacts of WASH on child mortality, which is weighted heavily in disability-adjusted life year (DALY) calculations (Cairncross and Valdmanis, 2006). Observational studies are needed to measure severe outcomes like mortality where withholding co-interventions (e.g., oral rehydration salts to treat severe diarrhoea) from control groups would be unethical. Observational studies are also needed to measure severe outcomes like mortality where withholding co-interventions (e.g., oral rehydration salts to treat severe diarrhoea) from control groups would be unethical. Observational studies are also needed to evaluate policy-relevant relationships between exposures, which are not amenable to researcher experimentation, and outcomes; for

example, the effect of diarrhoea episode duration on pneumonia (Schmidt et al., 2009). Doing so in a timely and rigorous way usually requires the use of natural experiments, using methods like RDD (Villar and Waddington, 2019). It is not clear to what extent these approaches are used effectively in WASH evaluation research.

1.6 Addressing bias in research

> "There exists a natural disinclination on the part of the head of a family to disclose intimate domestic details to others, and this added to a reasonable suspicion as to the motives behind the investigation and doubt as to the use which may be made of the information given, renders the collection of data a matter of difficulty."
>
> M'Gonigle and Kirby (1937, pp.193-4).

There are also concerns with the implementation of impact evaluation methods, including in WASH sector evaluation work, potentially causing biased effect estimates (Waddington et al., 2017). All quantitative causal studies are subject to a range of biases, relating to the design, implementation, and the wider relevance of the study (Shadish et al., 2002). For example, the well-conducted RCT is the preferred instrument of causal inference, but RCTs can have methodological problems in implementation such as contagion (contamination of controls), problems with the way randomisation was conducted, non-random attrition, and so on, causing bias (Higgins et al., 2011). Non-randomised studies are, however, potentially at higher risk of bias than their experimental counterparts (Sacks et al., 1982), perhaps the most critical for causal inference being confounding and biases in reporting (Higgins et al., 2012). They are also more difficult to assess, requiring greater qualitative appraisal than RCTs usually involving an understanding of theory. Hence there is a need for rigorous and transparent critical appraisal of these studies in research synthesis and policy research work (Waddington et al., 2017).

Much evidence from first generation evaluations measured efficacy rather than effectiveness, scalability or sustainability (Waddington et al., 2009). Problems with sustained adherence are well known in the household water treatment literature (e.g., Quick et al., 2002; Waddington et al., 2009).

Where interventions appeared effective (or ineffective) in reducing self-reported disease incidence, it was unclear if this was because compliance rates were high (or low), or because of unobserved confounding due to measurement error. The diarrhoeal disease measurement literature has long identified the recall period and definition of disease used, among others, as important sources of bias when diarrhoea is measured by reporting rather than observation (Blum and Feachem, 1983). Social desirability (courtesy) bias, where participant self-reporting is affected in response to being questioned, and survey effects (where being surveyed sensitises individuals to interventions, thus promoting uptake) have been shown to cause errors in open (unblinded) WASH impact studies using self-reported outcome measurement (Schmidt and Cairncross, 2009; Zwane et al., 2011).

Sometimes, the design of the interventions themselves is inappropriate. For example, three high-profile randomised controlled trials (RCTs) were conducted recently to assess the impact of WASH interventions on nutrition: WASH-Benefits in Bangladesh (Luby et al., 2018) and Kenya (Null et al., 2018) and Sanitation, Hygiene, Infant Nutrition Efficacy (SHINE) in Zimbabwe (Humphrey, 2019). The studies were not able to detect any effects on child linear growth, and only in Bangladesh was diarrhoea reduced. A consensus statement from Europe and the US has been published, challenging the efficacy of the WASH interventions in addressing faeco-oral pathogenic contamination in the contexts where they were implemented, and therefore the generalisability of the findings (Cumming et al., 2019).[13]

While the focus of this Thesis is primarily summative evaluation (counterfactual analysis), formative evaluation of process (factual analysis) is an important component in establishing effectiveness (White, 2009). Early WASH sector evaluation guidelines promoted the collection of process and intermediate outcomes (Cairncross et al., 1980; WHO, 1983). Evidence on processes may include implementation of fixed investment activities (e.g., hardware construction and community triggering) and recurrent service delivery activities (maintenance and follow-up). Intermediate outcomes relate to beneficiary knowledge, access to and uptake of interventions, user satisfaction and compliance or adherence. Data on adoption and adherence

[13] Ross (2019) gives an overview of the main arguments, including articulating why the incremental nature of the improvements made over baseline water and sanitation conditions (from 'close to basic' to 'basic' provision) was unlikely to lead to big reductions in communicable disease and malnutrition.

by beneficiaries in the context of theory-based impact evaluations help explain why the impacts have, or have not, occurred (Blum and Feachem, 1983; White, 2009; Waddington et al., 2009). In addition, adherence data can enable triangulation of findings for final outcomes when outcomes data are considered unreliable (Blum and Feachem, 1983), such as carer-reported morbidity in unblinded trials (Schmidt and Cairncross, 2009), or where they are measured in the context of uncontrolled longitudinal designs (Barreto et al., 2007). As indicated in Chapter 4 below, process information is needed to establish the risk of bias due to deviations from intended interventions, in impact evaluations.

It is therefore important to measure adherence and understand how it is affected by implementation. For example, the Minimum Evaluation Procedure (WHO, 1983) argued that evaluations should focus on the functioning of the facilities, and their use, which have greater diagnostic power to improve a programme than health impact evaluations. Different types of evaluation have different purposes (Figure 1.9). Mark and Lenz-Watson (2011, p.197) argued for "going beyond the bare-bones randomized experiment by (a) testing for possible moderated effects… (b) conducting mediational tests of possible mechanisms by which the treatment effect would occur, and/or (c) more generally, using multiple and mixed methods to complement the strengths and weaknesses of the randomized experiment." An emerging literature is now demonstrating the value of mixed-methods evaluation to answering these types of policy questions (Shaffer, 2013; Jimenez et al., 2018) including applications to WASH interventions (e.g., deWilde et al., 2008; Aunger and Curtis, 2016).

Figure 1.9 Purposes of two main types of evaluation

Formative	*Summative*
Diagnosis Internal, for ownership Mainly qualitative Purposive sample	Accountability External, for credibility Standardised, quantitative Representative sample

Source: Sandy Cairncross, pers. comm.

Decision-makers need access to rigorous evidence, appropriately interpreted, on the effects of WASH intervention mechanisms, in different contexts, for different types of programme participants. However, global policy decision-making should not draw on the results of single studies (or chosen groups of studies), but rather systematic reviews examining the

totality of evidence (e.g., Leach and Waddington, 2014). This is because even rigorous studies are only able to provide evidence on the extent to which WASH programmes can help overcome challenges and improve outcomes in the contexts in which they are implemented. There are important reasons why the applicability of findings of single studies to other contexts, or the transferability of interventions, may be limited. For example, the limited effect on nutrition of providing basic latrines found by WASH-Benefits in Bangladesh and Kenya may not be applicable in Indian contexts where the extent of open defaecation is much greater (Coffey and Spears, 2018). Furthermore, many single studies, including rigorous studies like RCTs and natural experiments, are subject to design or implementation flaws, and therefore may be at 'high risk of bias' in estimating the magnitude of the effect size. Single studies are usually underpowered to detect statistically precise changes when effect sizes are small, or for population sub-groups of interest as will increasingly be relevant under the SDG aims to reach the most disadvantaged groups to 'leave no one behind' (Waddington et al., 2018).

High quality systematic reviews, on the other hand, aim to collect, appraise and synthesise all the rigorous evidence relevant to a question, critically appraise and corroborate the findings from individual studies, as well as providing a steer to decision-makers about which findings are generalisable and which are more context-specific (Lavis, 2009; Higgins and Green, 2011; White and Waddington, 2012; Waddington et al., 2012). Approaches have been developed to reach conclusions about generalisability transparently, in particular grading of recommendations, assessment, development and evaluations (GRADE) (Guyatt et al., 2011).

The systematic review literature in WASH research is mature, unlike many other fields of international development (Waddington et al., 2012). After the first studies by Steve Esrey (Esrey et al., 1985, 1991), the standard practice has been for reviews of impact studies to use inverse-variance weighted meta-analysis to synthesise effect sizes across studies, from Curtis and Cairncross (2003) onwards. Statistical meta-analysis of effect sizes enables researchers to account for the magnitude of the treatment effect in individual studies, and its statistical power, in pooling data across studies (Glass, 1976; Smith and Glass, 1977). Other methods of synthesis based on 'vote-counting', or null-hypothesis significance testing, where studies are given a vote for whether the finding is statistically significantly different from zero or not,

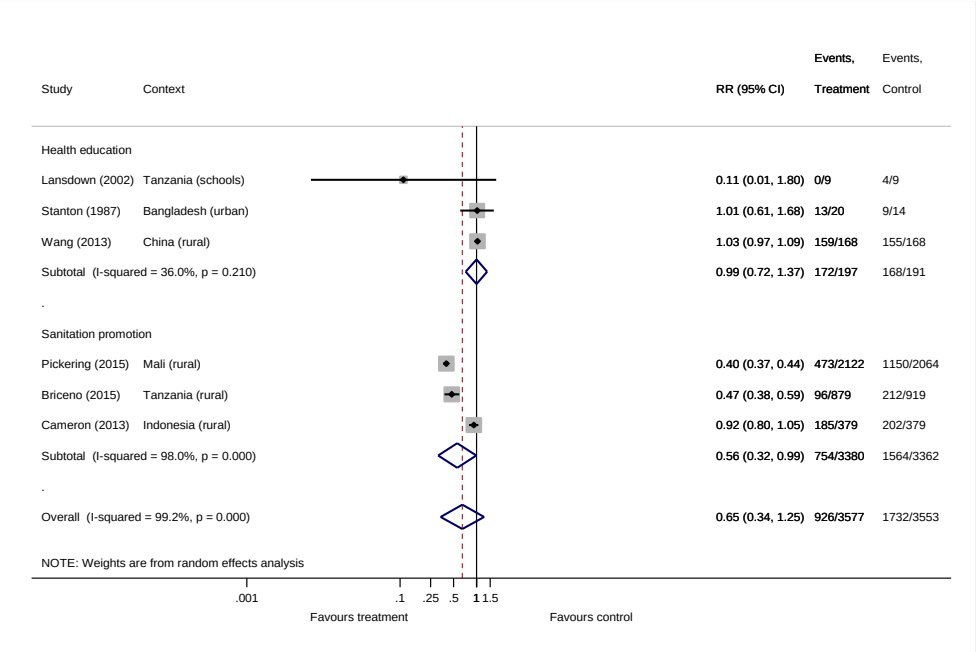
and weighted equally regardless of sample size, lead to biased conclusions (Cooper and Rosenthal, 1980).

The meta-analysis process has four distinct phases: calculation of standardised effect sizes from the studies (e.g., mean differences, odds ratios), critical appraisal of studies (risk-of-bias assessment), assessment of reporting biases (publication bias assessment), and synthesis including the possible statistical pooling across studies to estimate an average impact and explain heterogeneity in effect sizes. A final phase is to reach transparent conclusions about the generalisability of meta-analysis findings (Guyatt et al., 2011). These methods help overcome serious problems in interpreting evidence from single studies for decision-making (Waddington, 2014). Firstly, sample sizes in impact evaluations are often too small to detect statistically significant changes in outcomes, particularly when treatment effect sizes themselves are small, or if the study has not been powered to detect outcomes for sub-groups of interest like women and girls (Waddington et al., 2018), or for rarer outcomes like mortality (see Chapter 6). Meta-analysis takes advantage of the larger sample size from multiple evaluations and pools that evidence, exploring heterogeneity in findings statistically and graphically using forest plots (Higgins and Green, 2011).

Figure 1.10 gives an example of a forest plot showing effects on open defecation rates of hygiene education and sanitation promotion (CLTS). Studies are all open (unblinded) randomised controlled trials, evaluated using intention-to-treat (ITT). It is a good example of the importance of heterogeneity analysis, in this instance by moderator analysis of types of intervention mechanism. The pooled effect size across all studies does not indicate a significant reduction in open defaecation, measured with a high estimated level of statistical heterogeneity (I-Squared=99.5%). Moderator analysis by intervention type indicated that sanitation promotion caused an estimated 44 percent reduction in open defaecation on average (RR=0.56; 95 percent confidence interval (95%CI) =0.32, 0.99; I-squared=98%; evidence from 3 studies with 3,564 participants), whereas health education had no effect on open defaecation (RR=0.99; 95%CI=0.72, 1.37; I-squared=36%; 3 studies, 359 participants). However, there is residual unexplained heterogeneity. Sanitation promotion was less effective in Indonesia, possibly due to pre-existing latrine availability being higher, than in either African context. The authors of that study found significant impacts on open

defaecation for households that did not already have access to sanitation (Cameron et al., 2013). Following the hygiene education intervention conducted in schools in Tanzania, no open defaecation was observed (Lansdown et al., 2002). One can imagine it being more effective to change hygienic behaviour through simple messages among children in the controlled school environment, than it would be in the community. Indeed, factors of control are likely to be stronger in institutional settings, as also shown in a handwashing study conducted among U.S. Navy recruits who were instructed to wash hands five times a day and received "directive from the commanding officer that 'wet sinks' would be allowed to pass inspection (prior to this direction, recruit handwashing sinks were kept clean and dry in order to pass spot inspections)" (Ryan et al., 2001, p.80).

Figure 1.10 RCTs of interventions measuring open defaecation



Note: all outcomes were collected using self-report except Stanton (Stanton and Clemens, 1987) which used observation.
Source: author based on data reported in de Buck et al. (2017).

Secondly, all primary study literature is vulnerable to bias, which systematic reviews can help to overcome through critical appraisal. This is often done through assessment of risk of bias and generalisability (e.g., Waddington et al., 2012). For example, the review by Curtis and Cairncross (2003) included individual studies that estimated effects of handwashing on diarrhoea, shigella and typhoid that estimated null effects or even the opposite to those that would be predicted by theory. The ineffective handwashing studies were

also done in contexts where the water supply was limited, which would therefore limit participants' abilities to practice personal and domestic hygiene.

Thirdly, published studies are very unlikely to represent the full range of impacts that a programme might have. Publication bias (Rothstein et al., 2005), well-known across research fields, occurs where investigators are more likely to write up, and journal editors are more likely to accept, findings that can prove or disprove a theorem. Conversely, they are less likely to write or publish studies with null or statistically insignificant findings. Relatedly, investigators are more likely to undertake 'p-hacking' – that is, conduct multiple hypothesis tests, to identify statistically significant findings – which are the results that get reported in published papers. Meta-analysis can be used to identify publication biases resulting from small-study effects using formal statistical testing (e.g., Curtis and Cairncross, 2003; Fewtrell and Colford, 2004; Clasen et al., 2006; Waddington et al., 2009; Wolf et al., 2018).

Meta-analysis has been criticised by research and practice communities since its inception (e.g., Eysenck, 1978). Some of the concerns may be justified, such as those around pooling evidence from different contexts, without considering implementation factors, baseline conditions and methodological aspects of included studies (Wachter, 1988). However, meta-analyses of diarrhoeal disease commonly take baseline WASH conditions into account explicitly in the analysis, to allow effect sizes to vary by the incremental nature of the intervention over the control conditions (e.g., Fewtrell and Colford, 2004; Waddington et al., 2009; Hunter, 2009; Wolf et al., 2014, 2018). Where it is deemed inappropriate to pool findings across all studies – for example outcomes data are not collected consistently – narrative methods can be used to synthesise the evidence. A useful combined approach is to present evidence along the causal pathway (see Waddington et al., 2012; White et al., 2018).[14]

However, reviews on WASH topics have often focussed on summarising evidence about the efficacy of providing new or improved water supply, water

[14] For an example of a systematic review containing evidence along the full causal pathway, drawing on programme design, implementation and evaluation literature, see Waddington and White (2014) on participatory agricultural education.

treatment, sanitation and hygiene technologies to unserved populations, rather than effectiveness of WASH intervention mechanisms (e.g., behaviour change communication, subsidies and decentralisation) on uptake and use of improved WASH technology. Furthermore, a great number of the reviews that do exist focus on self-reported diarrhoeal morbidity outcomes, rather than a fuller range of socio-economic outcomes and health thought to be associated with improved WASH use. For example, nobody has investigated, critically appraised and synthesised the evidence on the impact of WASH interventions on childhood survival.

1.7 Structure of the Thesis

The Thesis presents the author's efforts to draw together rigorous evidence in four areas: measurement and evaluation of outcomes attributable to WASH programming; critical appraisal of statistical approaches to estimating the magnitude of the causal relationship between interventions, exposures and outcomes; and the scientific approach to the collection and synthesis of such studies to document the available evidence for making decisions about policy and programmes. Chapter 2 articulates the four Thesis Questions which the Thesis attempts to answer. Chapter 3 presents an evidence census for the WASH sector, drawing on existing and planned impact evaluations and systematic reviews, and examining their quality. Chapter 4 presents randomised and non-randomised evaluations of WASH interventions and develops a heuristic tool on which the probability of bias can be evaluated for different study designs. Chapter 5 analyses the biases in the literature and tests the relationship between predicted biases from the tool and the empirical evidence of bias, using systematic reviews of international development interventions. Chapter 6 presents results from a systematic meta-analysis of WASH impacts on child diarrhoea mortality. Chapter 7 concludes by articulating the extent to which the Thesis Questions have been answered, the limitations of the Thesis and its relevance for policy and future research.

Chapter 2 Thesis objectives

Impact evaluations and systematic reviews have been undertaken of WASH provision in L&MICs since the 1970s and 1980s, respectively, and are a rapidly growing area of WASH intervention research. There has been an explosion in the numbers of RCTs of WASH interventions. However, some types of programmes cannot be randomly assigned (e.g., universal programmes), some types of outcomes are measured with difficulty in prospective studies for ethical reasons (e.g., death in childhood), and some kinds of variables are not amenable to experimentation (e.g., exposures). There is still great interest in the findings of causal analysis in all these cases. There is also an interest in evaluating the impacts of existing programmes, which are designed by policymakers and assigned using methods other than randomisation, and estimation of long-term programme effects. It is therefore relevant to ask how prevalent these studies are.

It is also appropriate to ask whether the research resources devoted to impact studies and systematic reviews are relevant for those that the research is ultimately supposed to benefit. There are important concerns about the ways in which development research resources are distributed and the ways in which primary studies and evidence syntheses are routinely done. To take one example, many impact evaluations and reviews are done by researchers based at academic institutions in Western countries, and it is not clear to what extent researchers from L&MICs are involved substantively in these studies; not only is this unlikely to be a cost-effective approach in the long-term, but the research questions answered by these researchers may not reflect priorities of policy makers and poor people in L&MICs. Hence the first contribution of the Thesis is to analyse aspects of the political economy of WASH research in L&MICs.

Thesis Question 1: what types of interventions, outcomes and study designs can be, and are, covered in impact evaluations and systematic reviews of WASH interventions in L&MICs, and to what extent do the research resources devoted to impact studies and systematic reviews reflect the priorities of those that the research is ultimately supposed to benefit? The Thesis answers this first main question in Chapter 3, which presents a census

of impact evaluations, published in journals, books, working papers and organisational reports, conducted in low- and middle-income countries (L&MICs). The chapter contrasts the evidence and gaps identified with sector priorities, as expressed in the global burden of disease and a participatory poverty assessment. The chapter also examines the global distribution of WASH impact research production and the ethical and reporting practices that are common, by academic discipline, and the incentives provided by research funders and publishers in leading to the equilibrium in current research republishing practices.

Many studies on WASH topics measure diarrhoeal disease, the second biggest killer of children globally. It is beneficial to have agreement on, and common measurement of, key outcomes which are measured as routine across studies in a sector. However, most studies measure diarrhoea morbidity, which is assumed to be a good proxy for diarrhoea mortality, and there are important sources of bias affecting the reliability of reported illness in longitudinal studies, as well as other self-reported measures such as behavioural outcomes. The sources of bias in impact evaluations can be grouped into three domains: confounding and selection bias; bias in measurement of interventions and outcomes; and bias in analysis and reporting. While observational studies are more likely to be at risk of confounding bias than RCTs, they may be less subject to bias in measurement which results from participant expectations (e.g., Hawthorne effects).

There is, arguably, much greater scope for use of credible non-randomised approaches that theoretically have the benefits of RCTs (i.e., they can account for unmeasured confounding in attributing outcomes to WASH interventions) but can overcome some of the challenges in order to answer pressing questions for decision-makers. Some types of observational studies, called natural experiments (e.g., regression discontinuity designs), are able to estimate an unbiased causal effect in expectation without confounding, due to the way in which they are designed. However, most non-randomised studies (e.g., those using statistical matching and multiple regression), must rely on untestable assumptions to generate an unbiased causal effect estimate, by adjusting for confounding in analysis. Often the evaluation of natural experiments and non-randomised studies is complex. Critical appraisal, including risk-of-bias approaches used in systematic reviewing, has traditionally not taken this complexity into account adequately. Given

the large amount (US$ 100s of millions) of development funding dedicated to individual studies in the past decade or more, and the high profile that many prominent studies attract, it is appropriate to ask about the rigour and relevance of these studies.

Thesis Question 2: how can critical appraisal tools be operationalised to enable researchers to assess bias transparently and consistently for different types of quantitative causal study (including RCTs, natural experiments and other types of non-randomised study) and assess their relevance for decision-making? The Thesis answers this second question by further developing and piloting a tool to evaluate internal and external validity, applying it to a selection of WASH impact evaluation studies (Chapter 4).

Evidence synthesis collects, critically appraises and synthesises the results of multiple individual studies. Synthesis work can tell us about rigour and relevance of individual studies to the settings in which they have been conducted, and whether more generalisable lessons can be drawn to inform policy, programme design and delivery in many contexts. Evidence synthesis includes methods such as systematic review, meta-evaluation, statistical meta-analysis and realist synthesis, among others. The unifying feature of these approaches is their collation of multiple sources of evidence and the critical appraisal and synthesis of findings to answer questions about generalisability and context-specificity of evaluation findings.

Many systematic reviews and have been conducted to synthesise findings about the effectiveness of water, sanitation and hygiene technology provision, usually on diarrhoea morbidity using statistical meta-analysis. But it is not clear how useful they are in informing decision-making about particular WASH intervention mechanisms (e.g., community-led sanitation promotion) or ways of achieving particular outcomes in particular contexts. In addition, while the death of a child will be an important outcome for each household that has to face it, other health and socioeconomic outcomes are likely to be more important in determining acceptability, and therefore household demand for, new WASH technologies on a day-to-day basis. It is also possible that some WASH promotional interventions may not contribute to final quality of life outcomes, due to the long results chain and large number of other factors which influence outcomes of interest.

Thesis Question 3: to what extent are the biases, which are predicted in theory, borne out by empirical relationships between study effect estimates in practice? There is particular interest in evaluating whether non-randomised studies, including natural experiments, when well conducted, can produce the same effects as RCTs in practice. Chapter 5 aims to answer this question by analysing replication studies. The first section synthesises evidence from over 20 systematic reviews and meta-analyses of interventions across various international development topics (e.g., agriculture, climate change, economic development, education, governance). These reviews, which synthesise multiple external replications – that is, studies assessing the same or a similar intervention and outcome in different contexts and target populations – have used various iterations of the critical appraisal tool presented in Chapter 4. The analysis focuses on the relationship between predicted bias using the tool ('low risk', 'some concerns' and 'high risk of bias'), and the distribution of pooled effect sizes obtained from random effect meta-analysis.

The second part of Chapter 5 synthesises evidence from internal replication studies in international development – that is, studies that, for the same context and target population, compare the results of a benchmark study (usually a well-conducted RCT) with a NRS estimator. The purpose of this section is to validate the critical appraisal tool, to ensure it is based on empirical evidence about the relationship between probable bias and differences in study effects. Fixed-effect meta-analysis is used to synthesise that evidence.

Many systematic reviews have been conducted to synthesise findings from impact evaluations about the effectiveness of water, sanitation and hygiene technology provision on diarrhoeal illness in low- and middle-income countries (L&MICs). But the underlying assumption of these analyses is that diarrhoea morbidity is a good proxy for diarrhoea mortality, which is the biggest component of the global disease burden relating to inadequate WASH. There is no existing systematic review of child mortality data outcomes due to WASH, despite the large number of observational NRS estimating the relationship, as well as the presentation of child mortality in participant flow diagrams in trials.

Thesis Question 4: what are the effects of WASH provision on child mortality and do the effects vary by intervention and technology? Answering this fourth main question, considered in Chapter 6, is done through a comprehensive systematic review of evaluations assessing the impact of WASH on mortality. Data on the effects of WASH on mortality from studies in Chapter 3 are collected and critically appraised using the tool from Chapter 4, and synthesised using the greater statistical power of meta-analysis over single studies, nearly all of which were not powered to detect significant effects in mortality. Correlational analysis is also done of whether the findings from WASH evaluations are substantively affected when studies are categorised by intervention or are assessed as having various threats to validity.

This page is intentionally left blank.[15]

[15] It should be included in any updates of Wright et al. (2014).

Chapter 3 On rigour, relevance and representation in WASH impact evaluation

> "Let us engage with priority questions of most importance to policy-makers and poor people in developing countries, and so use evidence to improve policies, programmes and projects, spend development resources more effectively and so truly to improve lives."
>
> White (2013, p.47)

3.1 Introduction

A standard systematic review is often completed within 12-24 months (Waddington et al., 2018). Reviews can take a long time to produce findings, quickly becoming outdated in such a way that they fail to answer the questions on they were commissioned in a timely manner (Whitty, 2015). One way to speed up the process of knowledge translation from systematic searches is the evidence map. Evidence mapping is an approach to present the extent of evidence on a topic in a user-friendly format (Saran and White, 2018). Evidence mapping has proven incredibly popular with researchers and development organisations (Phillips et al., 2017). It is an attempt to democratise access to information on scientific studies, which are frequently collected in journal articles and technical reports that are physically or technically inaccessible to decision-makers, and to communicate that information in a format that is user-friendly.[16]

This chapter presents the results of a census of WASH impact evaluations and systematic reviews in L&MICs to answer Thesis Question 1: what types of interventions, outcomes and study designs can be, and are, covered in impact evaluations and systematic reviews of WASH interventions in L&MICs, and to what extent do research outputs reflect sector priorities?

[16] Indeed, one aspect of the user-friendliness of mapping, and a key rationale for developing the evidence mapping approach, is to provide a more efficient way of communicating primary research gaps than 'empty reviews'.

Emphasis is therefore given not just to mapping the evidence, but also to assessing whether WASH research is fulfilling its purpose to inform decision making. Section 3.2 presents the policy context that motivated this research. Section 3.3 presents inclusion decisions and the search. In Section 3.4, systematic reviews are discussed. Section 3.5 discusses WASH impact evaluations and examines how research priorities relate to priorities relevant for decision makers. Section 3.6 presents information about the quality of studies and whether reasonable ethical standards in research conduct are being met. The final section concludes.

3.2 Progress towards global targets and the need for greater efficiency in resource use

A number of strategic global initiatives have been established to monitor WASH sector activities and outcomes, to promote results-based management. Of particular note, the WHO/UNICEF Joint Monitoring Programme (JMP) provides data on access to and use of water and sanitation at country and regional levels since 1990.[17] JMP data, used extensively in this section, indicate great strides have undoubtedly been made in recent decades towards addressing global poverty and promoting access to and use of WASH services. The MDG water target was declared met at the global level (WHO/UNICEF, 2013). However, in 2017, the year pertaining to the latest global estimates, 144 million people still used surface drinking water directly from a river, pond, canal or stream, 435 million people used unprotected wells, springs or other unimproved sources, and 206 million used improved water that required more than 30 minutes roundtrip to collect.[18] There also remain big regional inequalities in access. In sub-Saharan Africa, 416 million people still use surface water, unimproved drinking water sources, or have limited access to improved services (requiring more than 30 minutes round-trip to collect). In South Asia, 137 million use surface water, unimproved water or have limited services, and in East Asia and the Pacific (EAP), 165 million people use them. The biggest improvements in access to drinking water have been in Asia, but coverage for 2.14 billion people in EAP and 1.65 billion in South Asia remains 'basic'. This means improved drinking water is

[17] The WHO and UN Water's Global Analysis and Assessment of Sanitation and Drinking-Water (GLAAS) monitors global activities (resource flows and policy commitments) biennially since 2008. UN Water also produces an annual synthesis report on progress in SDG6 (UN Water, 2018).
[18] WASH access and use data in this chapter are from https://washdata.org/.

provided at the community level or, if provided on premises, the supply is unreliable or contaminated (Table 1.1).

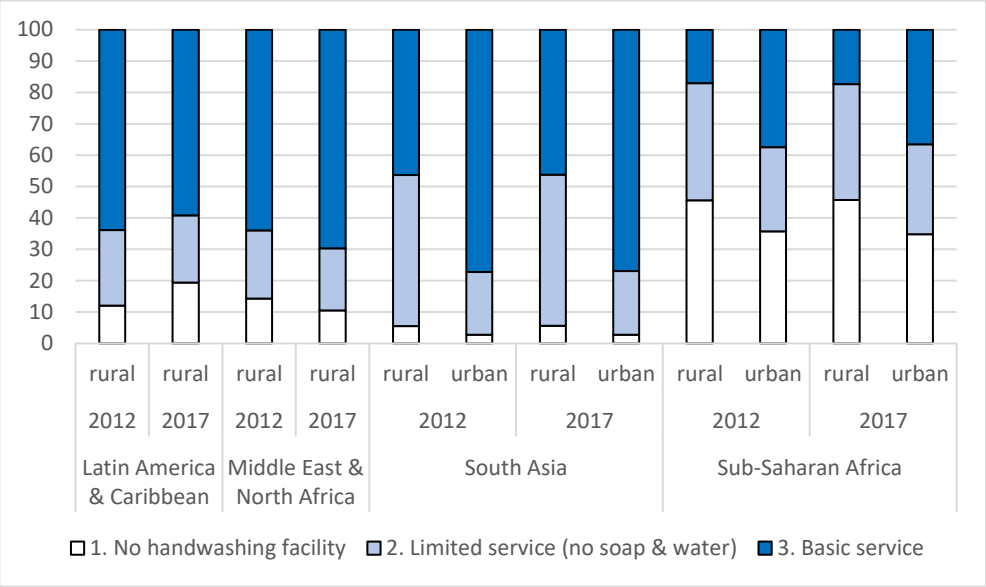
In 2008, at the MDG mid-point, recognising the more limited progress in improving access to and use of safe sanitation, the United Nations hosted the International Year of Sanitation. Unfortunately, the target for the MDG sanitation indicator, defined as the use of unshared, improved sanitation, was missed at the global level and in most countries in South Asia and sub-Saharan Africa by a wide margin (United Nations, 2015).[19] Of the 1.4 billion people who defecate in the open or use unimproved sanitation, 505 million are living in South Asia (of which 375 million are in India) and 546 million in sub-Saharan Africa. A further 620 million share limited sanitation facilities with two or more households (233 million in South Asia, 188 million in sub-Saharan Africa and 145 million in EAP). At the end of the MDG period in 2015, 4.5 billion people lacked access to safely managed sanitation, where excreta are disposed of safely in situ or offsite (UN Water, 2018).

Available data on access to hygiene facilities (Figure 3.1) indicate that the biggest share of people without access to even basic hygiene facilities is in sub-Saharan Africa and South Asia, where no improvements were made in 2012-2017. Over 80 percent of rural Africans, 530 million people, do not use a handwashing facility or use limited services without soap and water. Over half of those in rural South Asian, 640 million, also have no or limited handwashing services.

Rural households currently comprise the majority with inadequate facilities, although rapid population growth in urban areas means that urban access, particularly to sanitation and hygiene, is a growing policy issue (Bhatia and Falkenmark, 1993; WHO, 2018). In urban areas, 138 million people in South Asia and 267 million in sub-Saharan Africa lack access to basic handwashing. Data are not available on access to handwashing facilities in East Asia and the Pacific, or in urban areas of Latin America and the Caribbean (LAC) and Middle East and North Africa (MENA). Ensuring urban populations get access to adequate WASH services will become more important due to rapid population growth in these areas (United Nations, 2018).

[19] This relatively 'uneven progress' in reaching WASH sector targets was in part due to the sanitation indicator, defined as unshared by households, being harder to reach than the water indicator, which included shared facilities at the community level (Cumming et al., 2014).

Figure 3.1 Household hygiene access (% of population using service)



Note: data not available for EAP.

Source: data collected from https://washdata.org/.

The targets and indicators with direct relevance for WASH programming for consumption in households and public facilities are listed in Table 3.1. Reaching these targets will be challenging, and not just for sanitation and hygiene. For example, only 15 countries with less than 95 percent coverage are on track to achieve universal coverage of basic drinking water, only 14 countries with less than 95 percent coverage are on track for universal basic sanitation, and only 18 countries are on track to eliminate open defaecation (WHO/UNICEF, 2017).

Hutton and Varghese (2016) have estimated the capital cost of reaching those remaining unserved with basic water, sanitation and hygiene services at US$ 28 billion (2015 prices) per year from 2015-30, while the capital cost of providing safely managed services for all under SDG 6.1 and 6.2 is US$ 114 billion per year.[20] Most of the costs of WASH needs are borne by households and domestic government. A recent Global Analysis and Assessment of Sanitation and Drinking-Water (GLAAS) survey of 25 countries estimated 66 percent of financing for WASH was provided by households and 24 percent by government. In contrast, external financing through foreign aid (grants

[20] This comprises estimated capital costs of providing safe water at US$ 37.6 billion per year, basic sanitation at US$ 19.5 billion per year, safe faecal waste management at US$ 49 billion per year and hygiene at US$ 2 billion per year (Hutton and Varughese, 2016, p.7).

and concessional loans) comprised only 2 percent overall (WHO, 2017). However, aid inflows are a significant proportion of expenditure on WASH in many individual countries; in the same GLAAS survey, aid was the biggest non-household source in 18 countries out of 42, including Bangladesh and Cambodia in Asia, Cuba in Latin America, and Burundi, Kenya, Lesotho, Madagascar, Mali, Zambia and Zimbabwe in sub-Saharan Africa.

SDG target 6.A is to expand aid to domestic WASH budgets by 2030. Real aid disbursements to L&MICs to water and sanitation steadily increased in the past two decades, more than trebling to US$ 7.3 billion (2017 prices) between 2002 and 2016 (Figure 3.2). This was mainly due to increases in Development Assistance Committee (DAC) and multilateral donor disbursements, although emerging donors (non-DAC bilateral sources and private donors) are increasingly important sources. Aid disbursements follow commitments with a lag due of up to 10 years – that is, it is only after this period that disbursements reach levels previously committed. Total aid commitments to WASH fell in 2012-15, possibly because of limited absorptive capacity in the sector. Referring to this decline, the WHO and UN Water raised the concern that "the possibility of future reductions in aid disbursements does not align with global aspirations" (2017, p.ix). This concern appears to have been realised by the reduction in aid disbursements in 2017 to under US$ 7 billion due to multilateral disbursements. Aid commitments have risen above US$ 9 billion in 2017 and 2018,[21] but remain far below what is likely to be needed to make the SDGs achievable policy goals.

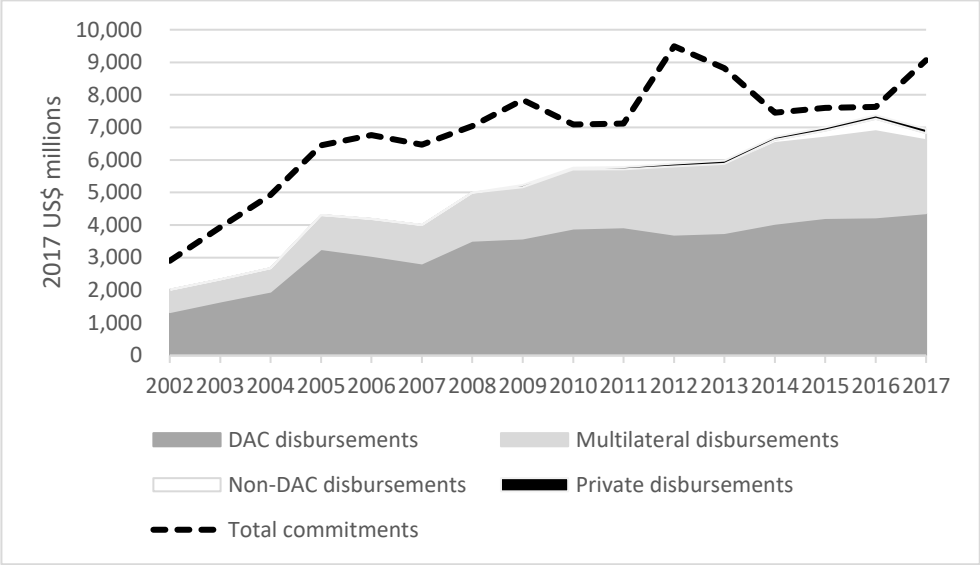
[21] Total commitments in 2018 were US$ 9.3 billion (https://stats.oecd.org/).

Table 3.1 SDGs relevant for WASH in households and public facilities

SDG	*Target definition*	*Indicator*
6.1	To provide safe and affordable drinking water for all by 2030.	Proportion of population using safely managed drinking water that is from an improved drinking water source, located on premises, available when needed and free from contamination.
6.2	To provide adequate and equitable sanitation for all and end open defaecation by 2030, ensuring that everyone has access to at least a basic toilet and safe waste disposal system, paying special attention to the needs of women, girls and vulnerable people.	Proportion of population using safely managed sanitation services, defined as an improved facility where excreta is treated and disposed of in situ or off-site.
6.2	Provide universal access to a basic handwashing facility with soap and water by 2030.	Proportion of population using a handwashing facility with soap and water.
6.3	Improve water quality by, among others, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally by 2030.	Proportion of wastewater safely treated and proportion of water bodies with good ambient water quality.
6.4	Substantially increase water-use efficiency and address water scarcity by 2030.	Freshwater withdrawal as a proportion of available freshwater resources.
6.A	Expand international cooperation and capacity-building support to developing countries in water- and sanitation-related activities and programmes by 2030, including water harvesting, desalination, water efficiency, wastewater treatment, recycling and reuse technologies.	Amount of water- and sanitation-related official development assistance that is part of a government-coordinated spending plan.
6.B	Support and strengthen participation of local communities in improving water and sanitation management.	Proportion of local administrative units with established and operational policies and procedures for participation of local communities in water and sanitation management.
1.4	To ensure all men and women, in particular the poor and vulnerable, have access to basic services by 2030.	Proportion of people living in households with access to basic services (including water, sanitation and hygiene).
3.3	End epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases (NTDs) and combat hepatitis, waterborne diseases and other communicable diseases by 2030.	Tuberculosis, malaria and hepatitis B incidence and number of people requiring interventions against NTDs.
3.9	To reduce substantially deaths and illnesses from hazardous chemicals and water pollution and contamination by 2030.	Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene.
4.A	Build and upgrade education facilities that are child, disability and gender sensitive and provide safe, non-violent, inclusive and effective learning environments for all.	Proportion of schools with, amongst others, basic drinking water, single-sex basic sanitation and basic handwashing facilities (as per the WASH indicator definitions).

Source: United Nations (undated).

Figure 3.2 Aid commitments and disbursements to WASH



Source: Creditor Reporting System https://stats.oecd.org/.

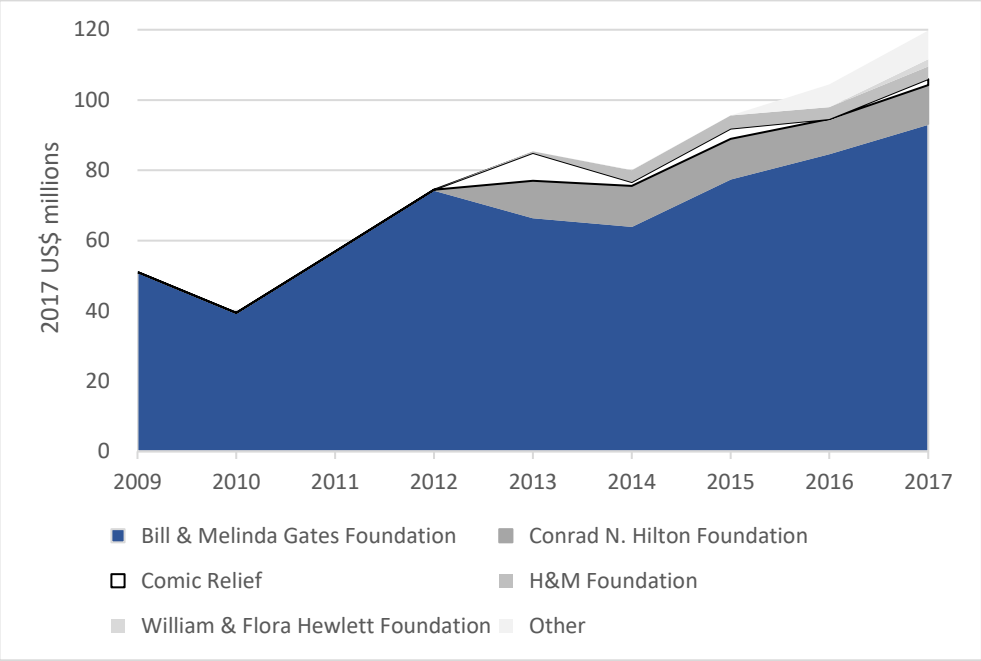
The ambitious targets, together with reductions in official development assistance, imply that big improvements in resource allocation are needed over a relatively short period of time. However, in the area of interventions faddism can easily propagate.[22] There is therefore increasing recognition of the role of rigorous evidence in facilitating efficiency improvements for meeting development targets (e.g., Waddington et al., 2018), by helping determine which interventions are appropriate for particular contexts in achieving desired outcomes. Private donors are of increasing importance to the generation of that evidence, by providing around one-third of aid disbursements to WASH research, comprising US$ 80 million or just over 1 percent of total aid to WASH.[23] The biggest by far is the Bill and Melinda Gates Foundation (The Gates Foundation), which gave US$ 93 million (2016 prices) to the sector in 2017 (Figure 3.3). The Gates Foundation is a major supporter of research and advocacy on effective and scalable interventions to improve sanitation demand.[24]

[22] For example, the Global Sanitation Fund of the Water Supply and Sanitation Collaborative Council (WSSCC), which was established in 2008, promoted the global scaling-up of CLTS in its activities before a single controlled evaluation had been conducted of the approach. While 21 evaluations of CLTS have been done since 2012, only around half measure health outcomes.

[23] Figure 3.3 shows aid channelled through "teaching institutions, research and thinktanks", which is a proxy for WASH policy research. This may underestimate total aid to WASH research since it does include aid through other channels which may undertake WASH research for example public sector, NGO and civil society, or multilaterals (e.g., aid to United Nations universities).

[24] Gates Foundation Water, Sanitation and Hygiene strategy, available at: https://www.gatesfoundation.org/what-we-do/global-growth-and-opportunity/water-sanitation-and-hygiene (accessed 18 February 2020).

Figure 3.3 Private donor disbursements to water and sanitation



Source: Creditor Reporting System https://stats.oecd.org/.

In 2016, the UN proclaimed 2018-2028 the International Decade for Action on Water for Sustainable Development.[25] To provide universal coverage, including appropriately serving the most disadvantaged people, it will be necessary to promote effective interventions for different groups, particularly disadvantaged groups who are most likely to be hidden from coverage, in the contexts in which they are used in private (household) and public realms (e.g., schools, health facilities, places of work, commerce and recreation, streets and fields). This goal of this chapter is to democratise access to information about intervention effectiveness in WASH. It presents a map of evidence from primary studies and systematic reviews on the effectiveness of interventions to improve the consumption of water, sanitation, and hygiene at home as well as in communities, schools, and health facilities in L&MICs.

3.3 Study inclusion and searches

Evidence maps are not a substitute for systematic reviews for two main reasons. Firstly, the standards of searching undertaken in evidence mapping are not usually as exhaustive as those for systematic reviews. For example,

[25] https://www.unwater.org/new-decade-water/ (accessed 17 November 2020).

sources may be limited to English language or by date; reference snowballing (citation tracing and bibliographic back-referencing) may not be undertaken. However, to produce this WASH evidence map, searches were done to the standards that would be taken in a 'high confidence' systematic review (Lewin et al., 2009), including searches for ongoing studies. The evidence map is therefore presented as an evidence census. Secondly, maps do not usually critically appraise or extract policy-relevant findings from primary studies. Chapter 6 presents a synthesis of evidence that draws on the studies collected here.

The census includes supply-side interventions to promote access to water, sanitation or hygiene services (e.g., direct provision, private sector involvement, capacity building), demand-side interventions promoting use of services (e.g., consumer behaviour change communication (BCC), consumer subsidies and microloans) and approaches addressing supply and demand (e.g., decentralised delivery through community-driven development, CDD). It also aims to go beyond 'diarrhoea reductionism' (Chambers and von Medeazza, 2014) by incorporating behaviour change (e.g., water treatment practices, open defecation, and time use), health (e.g., respiratory infections, enteric infections and mortality), nutrition and anthropometry (including enteropathy), and socioeconomic outcomes (e.g., education and income).

Table 3.2 summarises the criteria for inclusion of populations, intervention, comparators, outcomes and study designs (PICOS), as well as language and time frame, as specified further in Chirgwin et al. (2021). The census covered intervention mechanisms promoting WASH for household and personal consumption. It excluded interventions in food hygiene in the workplace such as a market (e.g., Sobel et al., 1998), methods to control faecal contamination by animals in the yard (e.g., Oberhelman et al., 2006), and vector control methods such as fly spraying (e.g., Chavasse et al., 1999; Emerson et al., 1999). Interventions primarily supporting farms or businesses such as dam construction (e.g., Duflo and Pande, 2007) were also excluded, as were interventions for groundwater or irrigation management (e.g., Meenakshi et al., 2013). Likewise, flood and drought management interventions and river, lake, coastal zone and wetlands management were omitted.

Studies were excluded where there was no clear intervention being provided, such as the association between shared versus private sanitation and diarrhoea (Baker et al., 2016) or access to water treatment kiosks (Sima et al., 2012). This criterion omitted studies focusing on important but uncommonly measured outcomes like musculoskeletal disorders (Geere et al., 2018), pre-term births and low birthweight (Olusanya and Ofovwe, 2010).

Table 3.2 Summary of inclusion criteria for WASH evidence census

Criteria	*Definition*
Populations	Human populations in low- and middle-income countries (L&MICs), as defined by the World Bank at the time the research was carried out, provided WASH in endemic conditions. Populations of any age, sex, gender, disability or socio-economic status were included. Populations in epidemics were excluded.
Interventions	Demand-side (behaviour change communication, subsidies, microloans, legal measures), supply-side (direct hardware provision, privatisation and nationalisation, small-scale independent provider involvement, improved operator performance), or combinations of demand- and/or supply-side (decentralisation). Technology and place of use: water supply, water quality, sanitation, and/or hygiene in the household, community, school or health facility.
Comparators	Impact evaluations where the comparison/control group receives no intervention (standard WASH access), a different WASH intervention, a double-blind placebo (e.g., non-functioning water filter), a single-blind (e.g., school textbooks), or a pipeline (waitlist).
Outcomes	Behaviour, health, and socioeconomic outcomes. Studies that only reported measures of knowledge or attitudes were excluded. Willingness-to-pay (WTP) was included where based on real purchase decisions.
Study design	Randomised controlled trials, prospective and retrospective non-randomised studies, natural experiments, and systematic reviews. For time use outcomes only: the above plus reflexive controls. For mortality outcomes only: the above plus case-control designs.
Language	Studies in English, French, Spanish and Portuguese. Studies in other languages were included where an English translation was available.
Time frame	No study was excluded based on date of publication.

Source: Chirgwin et al. (2021).

Co-interventions with a major non-WASH component were also excluded. This typically excluded deworming chemotherapy (e.g., Miguel and Kremer, 2004) and nutrition interventions (e.g., Humphrey et al., 2019), although any WASH-only arms without co-interventions in such studies were included (e.g., Luby et al., 2018; Null et al., 2018). Finally, studies, or components of

studies, that collected and analysed purely qualitative evidence were excluded. For example, in a controlled study of slum upgrading by Parikh and McRobie (2009) in Gujarat, India, women reported saving time and labour, and having fewer back problems, because of no longer having to carry buckets of water. However, the information was collected using qualitative interviews and presented in quotation.

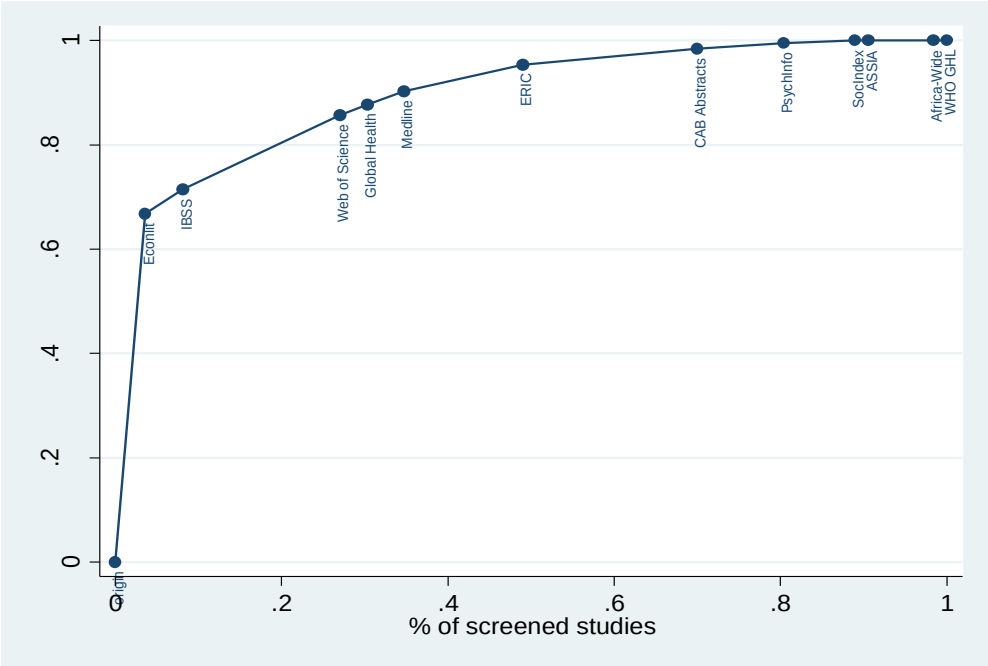
Systematic searches were done of both the published and 'grey' (i.e., non-peer reviewed) literature. A protocol, published in the Campbell library, details the search strategy (Waddington et al., 2018).[26] The Evidence for Policy and Practice Information Coordinating (EPPI) Centre's EPPI-reviewer 4 software was used to manage the screening process (Thomas et al., 2010). Once duplicates had been removed, there were 13,458 records for screening at title and abstract stage. To reduce resource requirements needed to screen this many studies at the title and abstract stage, machine learning was employed. The process of conducting systematic searches is becoming more and more demanding as more evidence is produced and more databases that require searching become available (Waddington et al., 2018). Hence, much of the time spent in conducting a systematic review is absorbed by the process of searching, screening and evaluating the available literature, often using word-recognition devices, with little time left for evaluating and synthesising the evidence. A large amount of researcher effort can be spared if we are willing to accept: a) that studies can be classified by a relevance

[26] The existing electronic database searches for an earlier evidence map and a 2017 systematic review (De Buck et al., 2017), were updated in March 2018. Searches were also run to cover the rest of the extended scope, particularly water behaviour change and health facility interventions. All search word lists were developed by an information retrieval expert and, in February 2018, eleven academic databases and four trial registry databases were searched. To capture grey literature, hand searches were conducted of key organisation websites. These included the Impact Evaluation Repository of the International Initiative for Impact Evaluation, the Asian Development Bank, African Development Bank, Inter-American Development Bank, Department for International Development, Improve International, International Reference Centre for Water and Sanitation (IRC-WASH), Oxfam, UNICEF, United States Agency for International Development, WaterAid, and the World Bank. Finally, the bibliographies of all included systematic reviews were checked to identify additional primary studies and systematic reviews. Reference lists of books, reports and evaluations were searched to identify additional WASH impact studies, particularly earlier ones that may not be captured in electronic searches (White et al., 1972; Saunders and Warford, 1976; Feachem et al., 1978; Cairncross et al., 1980; WHO, 1983; Khan et al., 1986; Briscoe et al., 1986; White and Gunnarson, 2008; Esteves Mills and Cumming, 2016). Finally, forward citation tracing searches were done in May 2020 for impact evaluations and systematic reviews that were identified as ongoing in 2018, and had since been completed.

score produced by a machine algorithm; and b) a reasonable margin of error in screening resulting in excluding some relevant studies.[27]

Figure 3.4 is an illustration of the potential for improvement. It shows the percentage of studies (vertical axis) as a function of the percentage of screened studies in each search database (horizontal axis) included in a recent review. The searches in the review were designed to be sensitive, meaning that they aimed to identify as many relevant studies as possible. The figure suggests that 20 percent of the searches delivered 80 percent of the studies included. It also suggests that, had the authors been willing to undertake searches with greater precision, omitting 20 percent of the evidence, they could have conducted the search in a fifth of the time.

Figure 3.4 Sensitivity and precision in systematic searches



Source: Masset (2020).

The problem with this example is that researchers do not know how many studies will be included and excluded from each database before conducting the search. The figure was calculated after the review was completed. However, clever methods are available to estimate the total population of studies.[28] For example, two early reviews of the effect of household water

[27] Reference snowballing may enable any studies missed by electronic searching to be identified.
[28] Method and original analysis proposed by Sandy Cairncross, pers. comm.

treatment on diarrhoea were incomplete: Fewtrell and Colford (2004) contained 13 studies, Gundry et al. (2004) contained 12, but only five studies were common to both reviews. By considering the two studies as a 'mark-release-recapture' experiment, this suggested a universe of 28 studies (95% confidence interval = 18, 88) which could be detected using an improved search strategy. A subsequent review conducted shortly after found 32 household water treatment studies (Clasen et al., 2006).

The method is due to Peterson and Lincoln (1930), defined in Krebs (2014) as:

$$\hat{n} = \frac{E_1 E_2}{S} \qquad (3.1)$$

where $\hat{n}$ is the estimated total population, E_1 and E_2 are the number of independent estimates by research teams *1* and *2*, and S is the number of observations in common. The formula produces an accurate estimate of the total number of available studies from two independent observations in expectation, because it is based on an identity. The number of estimates located by each independent research team, equal to probability $0<p<1$ of locating the total number of studies, is $p_1 n$ and $p_2 n$ respectively. One would also expect the independent research teams to find $p_1 p_2 n = S$ studies in common. Therefore:

$$\frac{E_1 E_2}{S} = \frac{p_1 p_2 n^2}{p_1 p_2 n} = n \qquad (3.2)$$

The method is biased in small samples. The corrected population size for small samples, defined as $E_1+E_2 \leq n$ and $S < 7$ (Krebs, 2014, Chapter 2, p.25), is estimated as:

$$\hat{n} = \frac{E_1(E_2 + 1)}{S + 1} \qquad (3.3)$$

which is unbiased for independent samples with replacement. The lower and upper limits of the 95 percent confidence interval (95%CI) for small samples is given as (Krebs, 2014):

$$\hat{n}^{upper}_{\hat{n}^{lower}} = \frac{1}{CI} E_1 \text{ where } CI = \frac{S}{E_2} \pm \left\{ 1.96 \sqrt{\frac{\left(1 - \frac{S}{E_1}\right)\frac{S}{E_2}\left(1 - \frac{S}{E_2}\right)}{E_2 - 1}} + \frac{1}{2E_2} \right\} \quad (3.4)$$

There are of course many reasons why systematic reviews on water, sanitation and/or hygiene might include different studies, or not be undertaken based on independent searches. Most obviously, included interventions or primary outcomes may differ. For example, many reviews have been restricted to health impacts like diarrhoea (e.g., Waddington et al., 2009; Clasen et al., 2015; Wolf et al., 2018), while a few others focus primarily on behavioural outcomes (e.g., de Buck et al., 2017; Garn et al., 2017). Or study design inclusion criteria may differ, with some restricting inclusion to studies evaluating a particular intervention (e.g., Clasen et al., 2015; Wolf et al., 2018) and others including exposures as well (e.g., Curtis and Cairncross, 2003; Waddington et al., 2009; Heijnen et al., 2014). In addition, there is a growing tradition of updating systematic reviews for new studies, so searches are not independent. Most recently, the systematic review of WASH and diarrhoeal morbidity by Wolf et al. (2018) updated searches and analysis done by Wolf et al. (2014), which itself was designed based on comprehensive reviews on the same topic by Waddington et al. (2009) and Cairncross et al. (2010). Waddington et al. (2009) was in turn an explicit update of Fewtrell and Colford (2004), which itself updated Esrey et al. (1985, 1991). Cairncross et al. (2010) originated from Curtis and Cairncross (2003) and Clasen et al. (2006).

Two recent reviews that did systematically search for the same intervention and outcomes – evaluations of the effect of sanitation promotion on behaviour change – are de Buck et al. (2017) and Garn et al. (2017). As far as it is possible to tell, these reviews were done independently, as neither cites the other.[29] Thirty-seven sanitation promotion studies were contained in the two reviews, of which only nine were are common to both. De Buck et al. (2017) included 18 studies, while Garn et al. (2017) included 28. Part of the reason for the difference is that Garn et al. (2017) were more inclusive on design, including, in addition to contemporaneously controlled

[29] Neither final report nor protocol (if available) were cited by either study team. A systematic review of child faeces disposal interventions, covering some of the same included studies as de Buck et al. (2017), was completed recently (Majorin et al., 2019). These reviews also appear to have been done independently, as neither study cites the other.

evaluations, reflexive controls (pre-test and post-test only). Applying equation (2.3) gives an estimated 55 studies in total (95%CI = 39, 101). Once again, this estimate is remarkably accurate: the searches undertaken for the evidence census found 53 studies of sanitation behaviour change.[30] This suggests there may be value in applying this method in analysis of bias in searches, and potentially other systematic review error checking (e.g., multiple coder verification).[31]

A related question is whether machines can support researchers in improving the precision with which searches are done. Much research and several projects are underway that employ machine learning algorithms to assist researchers in conducting systematic reviews (O'Mara-Eves et al., 2015; Tsafnat et al., 2014). In these trials, researchers screen a subset of the population of studies. The result of the screening process is fed into a machine which develops a rule to include or exclude a given study based on the information provided by the researchers. This is normally performed by a logistic regression where the dependent variable is the inclusion-exclusion of the study and the explanatory variables are words and combinations of words in the studies reviewed. The inclusion rule is then applied to a new subset of the data and the selection performed by the computer algorithm is returned to the researchers. The researchers at this point can perform an additional screening on the results of the search conducted by the computer, that can be fed back again to the machine to improve and refine the inclusion process at successive trials. In this way, the machine iteratively learns to include the studies using the criteria followed by the researchers.

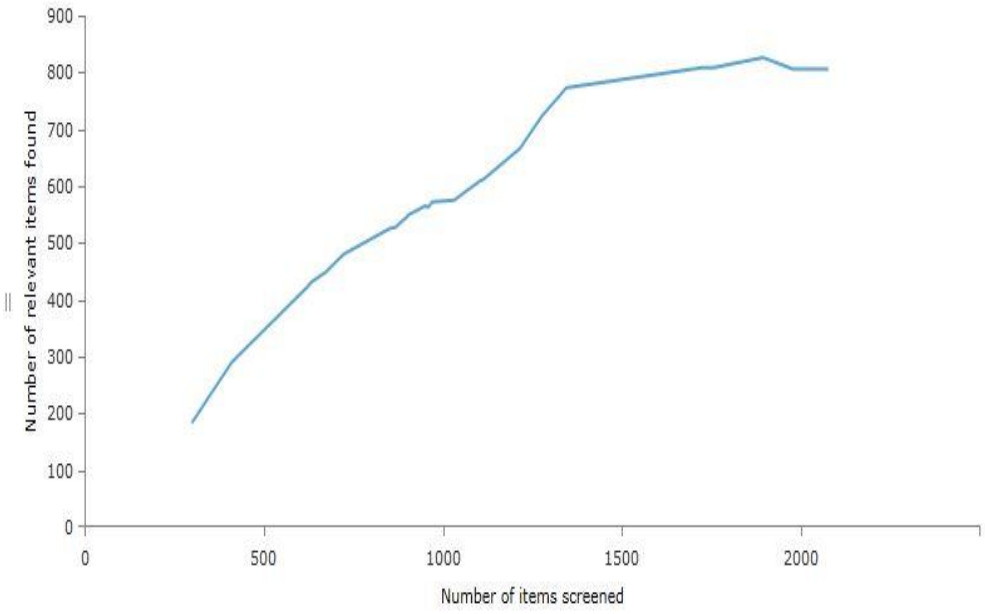
The machine learning software, which is integrated into EPPI-Reviewer, functions by identifying key words, through text mining, in included and excluded records. It then ranks studies from most to least likely to be included. This can be updated at regular intervals to reflect more recent inclusion decisions. Other studies looking at the effectiveness of this

[30] Sixteen studies featured in neither Buck et al. (2017) nor Garn et al. (2017), although five of these were published in 2017, presumably after the searches in those reviews had been completed. In addition to independence of sampling, an assumption of the method presented here is fixed population size. Methods for estimating populations of increasing size are shown in Krebs (2014).

[31] Due to restrictions on study design, only 34 studies were eventually included. The method is also accurate when applied to study arms: n = 32 x 46/23 = 64 (95%CI = 54, 78) estimated total study arms. Searches found 71 intervention arms, of which 52 were eligible for inclusion.

software found that it can often save up to 70 per cent of the workload with a loss of only 5 percent of the includable studies (O'Mara-Eves et al., 2015). After removing duplicates, two authors screened the records at the title and abstract stage until they did not find a single includable study for 100 consecutive records (Figure 3.5). A random sample of 100 of the remaining studies was then used to increase confidence that no studies had been missed. Ultimately, only 1,798 records were manually screened, a workload saving of almost 90 percent. Two authors then screened the remaining papers at full text.

Figure 3.5 Application of machine learning in WASH searches



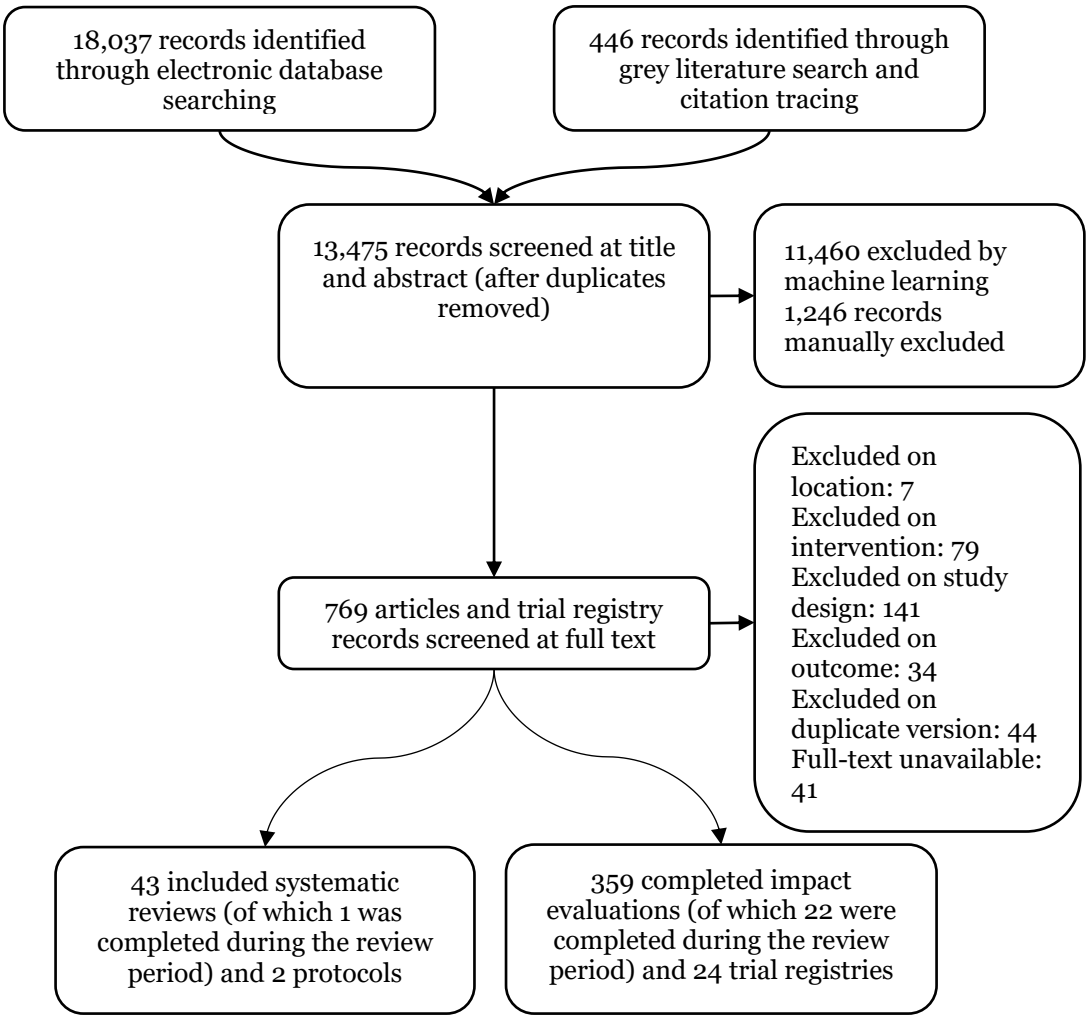
Note: the negative gradient in the curve at the 1,900 studies screened point was due to the decision taken to deviate from protocol by excluding non-WASH co-intervention studies and trial arms.
Source: EPPI-reviewer 4 (Thomas et al., 2010).

3.4 Findings about the quantity of completed and ongoing studies

The search results indicate that in total there are at least 358 completed and 22 on-going impact evaluations of WASH interventions in L&MICs, nearly three-quarters of which have been completed since 2008. There are also at least 43 systematic reviews and 2 protocols, of which all but four were completed after 2008. Figure 3.6 presents the preferred reporting items for

systematic reviews and meta-analyses (PRISMA) study search flow diagram.[32]

Figure 3.6 PRISMA study search flow diagram for WASH evidence census



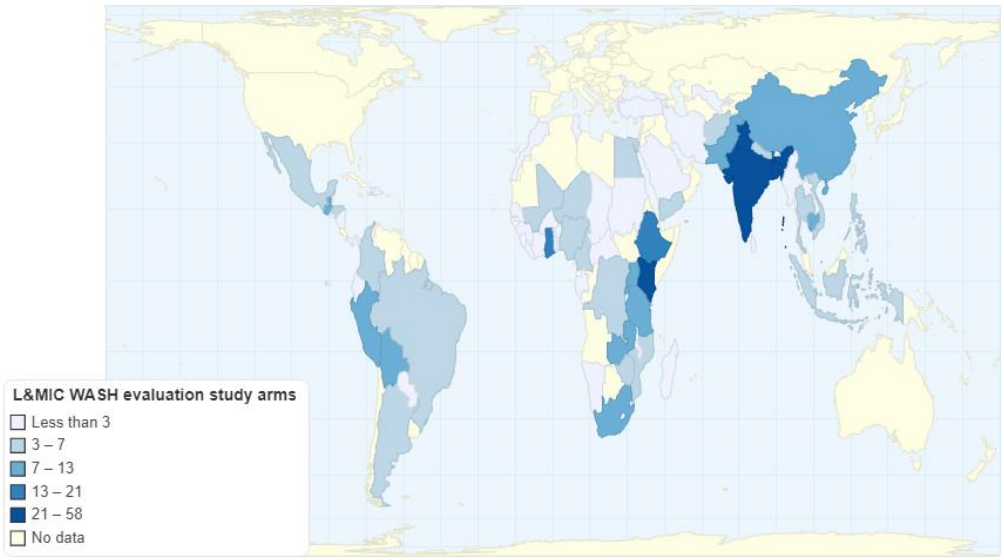
Source: Chirgwin et al. (2021).

[32] At the time the searches were completed in 2018, there were 336 completed and 46 on-going impact evaluations using quantitative counterfactual methods in L&MICs. There were also 42 completed systematic reviews of effects and three protocols. By May 2020, one systematic review (Majorin et al., 2019) and twenty impact evaluations had been published of ongoing studies (Acey et al., 2018; Arman et al., 2020; Augsburg et al., 2019; Batmunkh et al., 2019; Chauhan et al.; Cocciolo et al., 2020; Delea et al., 2020; Dreibelbis et al., 2018; Dupas et al., 2017; Friedrich et al., 2019; Gray et al., 2019; McGuinness et al., 2020; Kirby et al., 2019; Peletz et al., 2019; Rabbani, 2017; Reese et al., 2019; Trent et al., 2018; Vijayaraghavan et al., 2018; Viswanathan et al., 2019; World Bank, 2017).

3.4.1 *WASH impact evaluations*

Impact evaluations of WASH interventions have been conducted in 83 low- and middle-income countries (Figure 3.7). There is a high concentration of studies in Bangladesh, Kenya and India, each having over 50 WASH intervention study arms. In addition, Bolivia, Cambodia, Ethiopia, Ghana, Pakistan, Rwanda, and Uganda each have 10 or more.

Figure 3.7 Map of WASH impact evaluation interventions in L&MICs



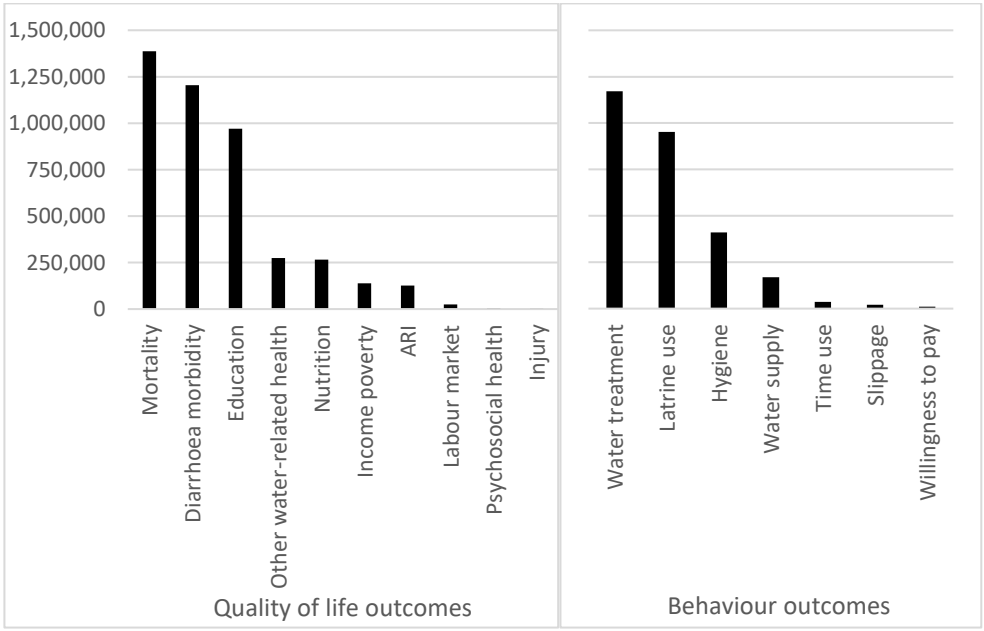
Source: created using chartsbin.com.

The total population included in WASH impact evaluations in L&MICs is at least 5 million participants. More than a million people have taken part in trials measuring, or had data collected on, child mortality and diarrhoea morbidity, and nearly a million have taken part in studies measuring education outcomes (Figure 3.8). Similarly, around a million people have participated in studies where water treatment and latrine use (including open defaecation) outcomes were collected. At the same time, however, very few have participated in studies measuring time use and labour market outcomes, willingness-to-pay in real-world scenarios, or studies measuring psychosocial health, injury.

Figure 3.9 plots the evolution of studies over time, indicating the marked increase after the International Year of Sanitation. Well over half of the studies (comprising over 250 trial arms) used randomised assignment (RCTs), indicating the extent of support in academic and research funding communities for this research method. Some RCTs have taken full advantage

of the power of the methodology by conducting comparative designs with prospective randomised assignment to alternate intervention mechanisms. Guiteras et al. (2015b) provided an example in Bangladesh comparing the effects of community sanitation promotion (CLTS) with subsidies on open defaecation.

Figure 3.8 Number of impact evaluation study participants by outcome

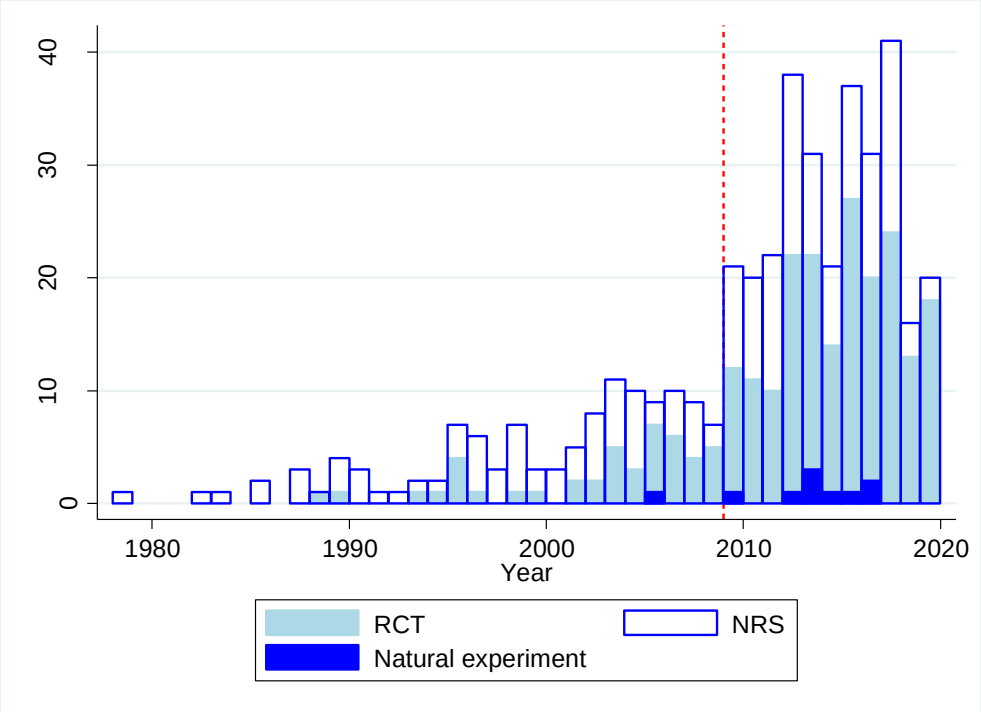


Typical non-randomised study designs include cross-section studies with statistical matching (e.g., Abou-Ali et al., 2009), group level panel data studies analysed at aggregated administrative levels (e.g., Galiani et al., 2005), individual-level panel data studies (e.g., Galiani et al., 2009), pseudo-panels with repeated cross-section from the same clusters (Galdo and Briceño, 2005), case-control studies (e.g., Meddings et al., 2004), prospective cohort studies (e.g., Shiffman et al., 1978) and pipeline studies (e.g., Cairncross and Cliff, 1987). In non-randomised studies using matching, the matching was usually done using statistical methods, although a few used 'naïve' matching, where observationally similar groups are compared without formal statistical tests (e.g., World Bank, 1998).

A small number of non-randomised studies (11) have taken advantage of existing data to conduct rigorous, and potentially highly cost-effective, evaluations with selection on unobservables, here called natural experiments (Ao, 2016; Calzada et al., 2013; Galiani et al., 2005; Galiani et al., 2009; Granados et al., 2014; Kosec et al., 2013; Spears, 2013; Tiwari et al., 2017;

Ziegelhoefer, 2012). Methods used to analyse data in natural experimental frameworks include regression discontinuity designs (RDDs) (Ziegelhoefer, 2012), interrupted time-series (ITS) (e.g., Duflo et al., 2015) and panel data regression (e.g., Galiani et al., 2005).

Figure 3.9 Total number of study arms by study design



Notes: dotted line shows the end of the International Year of Sanitation (2008). The apparent decline in production of studies post-2018 reflects the limited searches done in this map after 2018.

Prior to the International Year of Sanitation, the priority for intervention research had been efficacy studies of WASH technology provision, particularly of household water treatment and hand hygiene. For these 'first generation' impact evaluations, household water treatment interventions were the most studied technologies, and remain so (around 30 per cent) (Figure 3.10). However, more studies (e.g., Brown et al., 2012; Klasen et al., 2012) including two randomised encouragement trials (Devoto et al., 2010; Ben Yishay et al., 2017), broaden the evidence base on health impacts of water supply provision. The number of sanitation technology study arms has increased from 8 to 62, and there are similar magnitudes of increase of study arms examining hygiene (from 23 to 97).

These developments coincided with a shift, over the last 15 years, towards evaluation of WASH promotion, in 'second generation' impact evaluations of behaviour change communication using approaches like psychosocial 'triggering' (Figure 3.11). In sanitation, this is most commonly community-led total sanitation (CLTS). Hygiene promotion includes approaches like 'super-Amma' (super-Mum), which used the emotional driver nurture (the desire for a happy child) to incentivise improved handwashing practices (Biran et al., 2014).

Figure 3.10 WASH technologies by publication date

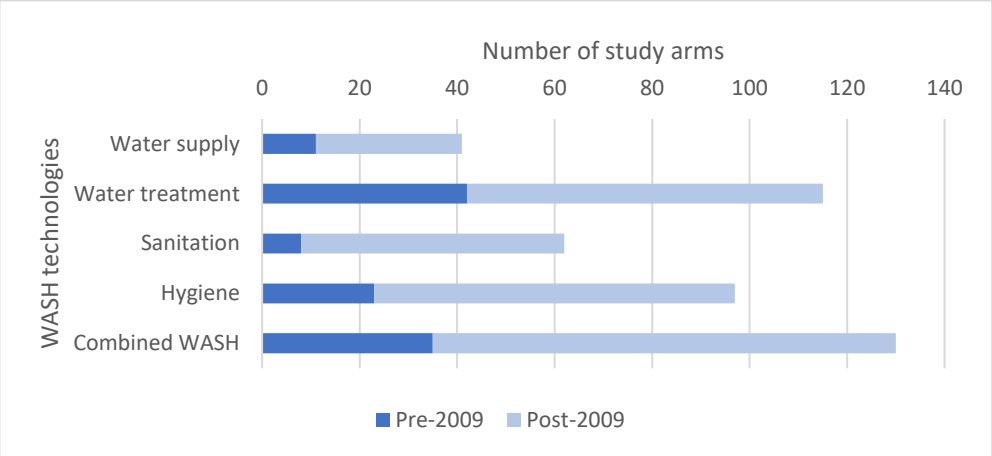
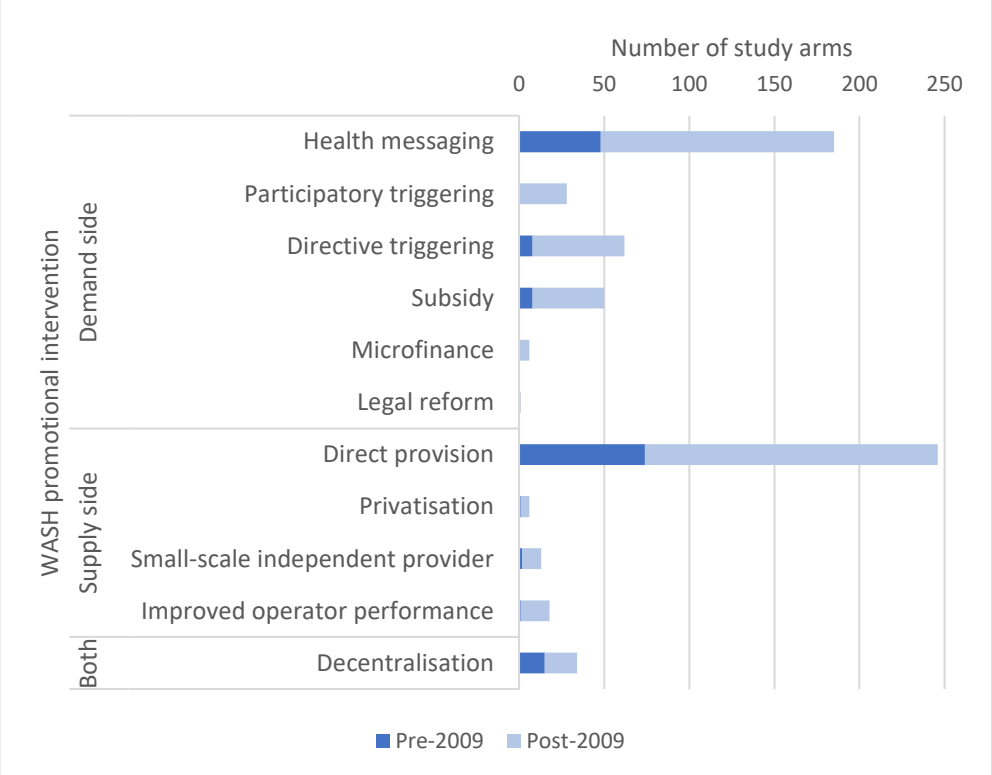


Figure 3.11 WASH interventions by publication date



Carer-reported child diarrhoea morbidity is the standard outcome measure used in WASH sector evaluations, and accordingly is by far the most reported outcome. There has been some increase in the number of studies looking at outcomes (e.g., time use) and interventions that disproportionately affect women and girls, but most studies still do not even report sex disaggregated outcomes, presumably due to low statistical power. Few prospective studies can assess child mortality as a primary outcome due to power and ethical reasons. Twenty-seven intervention studies have examined impacts of water provision and sanitation on child survival in L&MICs. These include Latin American studies conducted in Argentina (Galiani et al., 2005), Bolivia (Newman et al., 2002), Brazil (Rasella, 2013), Colombia (Granados and Sańchez, 2014), Ecuador (Galdo and Briceño, 2005), Honduras (Instituto Apoyo, 2000), Mexico (Venkataramani et al., 2013) and Paraguay (World Bank, 1998). Studies have also been done in South Asia – Afghanistan (Meddings et al., 2004), Bangladesh (Luby et al., 2018), India (Clasen et al., 2014; Spears, 2013), Nepal (Rhee et al., 2008), Pakistan (Bowen et al., 2012) – and others in Africa – Côte d'Ivoire (Messou et al., 1997), Egypt (Abou-Ali et al., 2009), Ethiopia (Gebre et al., 2011), Kenya (Crump et al., 2005; Null et al., 2018) and Mali (Pickering et al., 2015). Prospective studies examining child mortality are limited for ethical reasons required to measure death accurately, such as the need to withhold curative treatment – oral rehydration or clinical treatment. However, some prospective studies reported diarrhoea mortality (Messou et al., 1997; Luby et al., 2004; Bowen et al., 2012; Pickering et al., 2015) and it is possible to obtain all-cause mortality estimates from participant flow diagrams that should be commonly reported in RCTs (e.g., Bowen et al., 2012; Clasen et al., 2014; Luby et al., 2018; Null et al., 2018), explored further in Chapter 6.

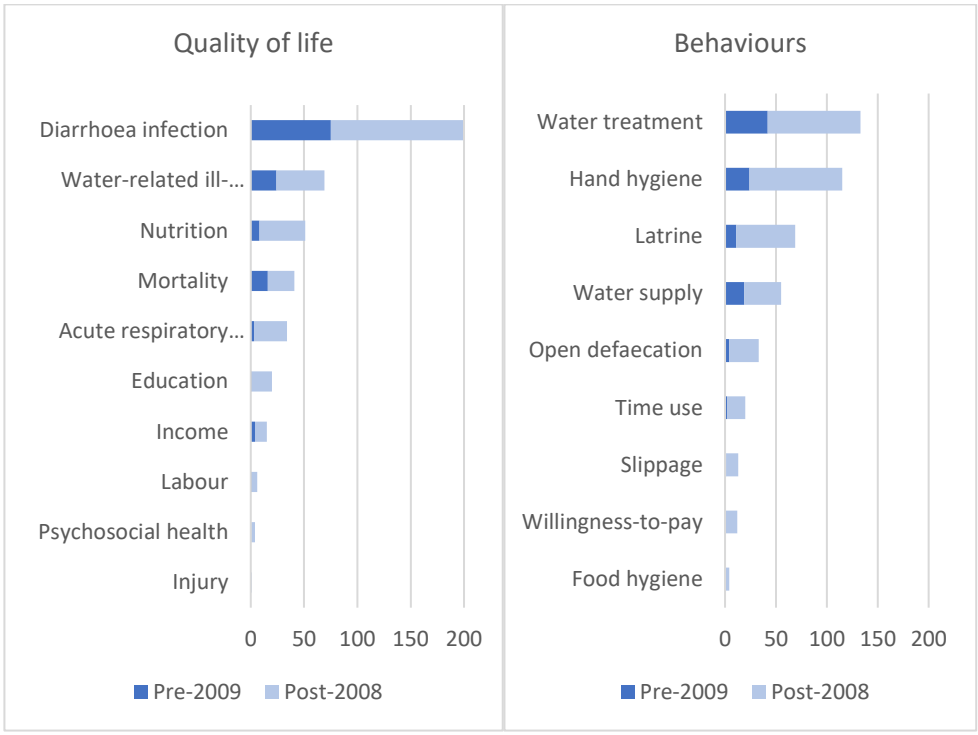
A systematic review from 2009 estimated that 71 completed study arms of WASH projects had been conducted measuring diarrhoeal morbidity (Waddington et al., 2009). The most recent systematic review of WASH and diarrhoea morbidity (Wolf et al., 2018) included 135 studies, and the evidence census presented here includes 186 study arms measuring diarrhoea morbidity, 119 of which were in studies published since 2008. More than a million people have taken part in trials measuring, or had data collected on, child mortality and diarrhoea morbidity.

Recognising the importance of WASH for controlling acute respiratory infections, the coverage of studies examining impacts on ARIs of hygiene promotion has also increased, with 35 study arms measuring acute respiratory infections including 31 in studies published post-2008, including large-scale studies in Vietnam (Chase et al., 2012), Colombia (Correa et al., 2012), Bangladesh (Huda et al., 2012), Guatemala (Arnold et al., 2012) and Egypt (Talaat et al., 2011). Studies of transmission of causative agents in unhygienic environments in L&MICs include acute respiratory infection like coronavirus (e.g., Esrey et al., 1988). However, given the importance of ARIs in the global burden of disease, and their enhanced importance in the coronavirus pandemic, the total of number of participants in WASH studies of ARIs, at only 125,000 in L&MICs, remains extremely limited (Howard et al., 2020).

In line with the other changes, there has been a shift in the commonly reported outcomes, including an increase in studies reporting behavioural outcomes (Figure 3.12). This is an important shift as the principal argument used by proponents of alternative delivery mechanisms is that they are more effective at changing these behaviours and therefore improving lives (e.g., Kar and Chambers, 2008). In addition, it is argued, interventions fostering marginal improvements in WASH behaviour may not cause sufficient changes at community level to improve quality of life outcomes like child nutrition or diarrhoea mortality (Geruso and Spears, 2018). However, very few studies measure sustainability of uptake or slippage back to old practices such as open defaecation, despite its importance for sustaining health improvements.

Nearly a million people have taken part in studies measuring education outcomes. Similarly, around a million have participated in studies where water treatment and latrine use (including open defaecation) outcomes were collected. At the same time, however, very few have participated in studies measuring time use and labour market outcomes, willingness-to-pay in real-world scenarios, or studies measuring psychosocial health, injury. And evidence of longer-term behaviours, including slippage back to bad practices, is extremely limited.

Figure 3.12 Number of impact evaluations by outcomes



The most frequently reported behaviours are handwashing, water treatment and handling, and latrine use. Many of the studies reporting hygiene behaviour, include measures of personal food hygiene; nearly 50 study arms specifically collect data on handwashing before food preparation, five report on the microbial contamination of food or eating utensils, and 17 report on other food hygiene outcomes, such as whether food is stored properly, and dishes washed appropriately. It is important that hygiene studies examine food hygiene outcomes, given the importance of food in faecal-oral disease transmission (Wagner and Lanoix, 1957). Studies collecting water supply behaviour outcomes include 40 study arms of interventions to reduce faecal contamination and six in Bangladesh of chemical contamination due to arsenic. There has also been an increase in the reporting of social and economic impacts. This is principally driven by a large increase in the number of studies reporting measures of education and cognitive development, and reflects the increase of studies being conducted in schools.

3.4.2 *WASH systematic reviews*

Systematic reviews of WASH studies include evidence from all global regions and cover a breadth of WASH technologies (that is, hardware and software, outcomes and, increasingly, promotional interventions. An estimated 43

completed systematic reviews have synthesised the findings of WASH provision (Figure 3.13). As impact evaluations make up the underlying body of research, systematic reviews predominantly focus on health outcomes, particularly diarrhoea and enteric infections.

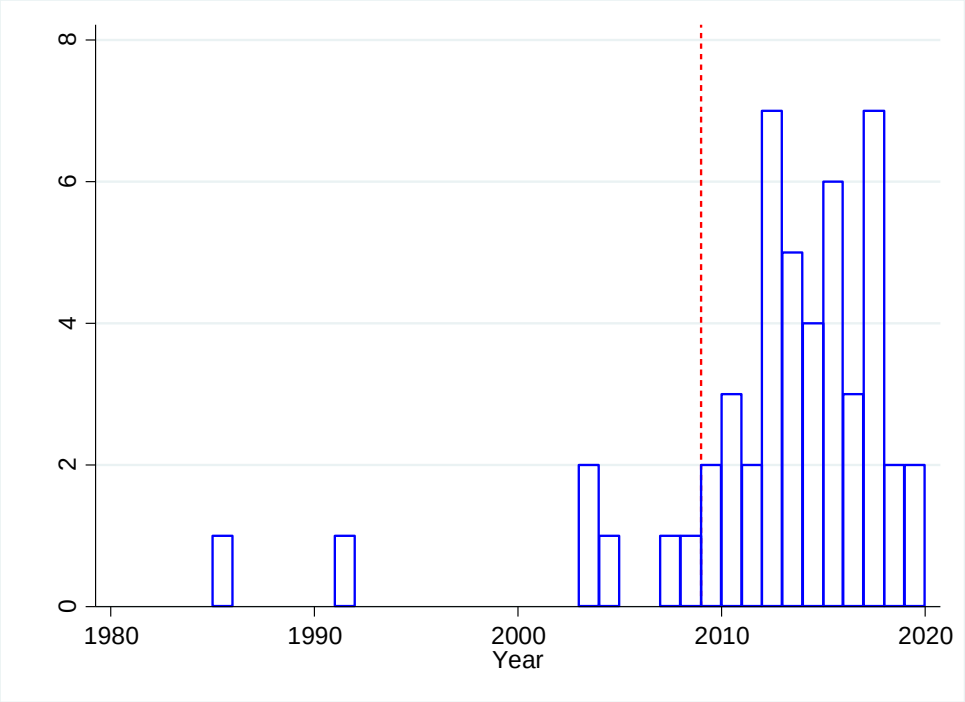
The classic systematic review, produced when systematic reviews had not yet been properly defined, was a series on the control of diarrhoeal disease in young children commissioned by the WHO Diarrhoeal Diseases Control Programme.[33] This included reviews of enteric infections associated with water and sanitation provision including diarrhoea (Esrey et al., 1985) and water-related infections (Esrey et al., 1991). Both reviews were explicitly restricted to published literature. Even so, Esrey et al. (1991) found large numbers of eligible studies (144 studies), due to comprehensive inclusion of outcome categories (diarrhoea, ascariasis, Guinea worm infection, hookworm infection, schistosomiasis and trachoma), and inclusivity by study design. Many 'first generation' reviews were subsequently done on diarrhoea morbidity (Curtis and Cairncross, 2003; Fewtrell and Colford, 2004; Clasen et al., 2006; Waddington et al., 2009; Clasen et al., 2010; Cairncross et al., 2010; Norman et al., 2010; Wolf et al., 2014; Clasen et al., 2015; Ejemot-Nwadiaro et al., 2015; Wolf et al., 2018). An increasing number of reviews are measuring other commonly evaluated outcomes, including 'neglected tropical diseases' such as helminth infections (Esrey et al., 1991; Ziegelbauer et al., 2012; Strunz et al., 2014; Freeman et al., 2017), trachoma (Esrey et al., 1991; Rabiu et al., 2012, Stocks et al., 2014; Ejere et al., 2015; Freeman et al., 2017), and Guinea worm infection (Esrey et al., 1991). Reviews have also been done of impacts of WASH on nutrition (Dangour et al., 2013), of WASH in schools (Freeman et al., 2014), and methods to reduce arsenic poisoning by contaminated ground water (Jones-Hughes et al., 2013).

A systematic review will be most relevant when the methodology is applied to a clearly defined research question, and preferably where eligible evidence is known about *a priori*. A common approach used in WASH systematic review and meta-analysis is to ask a question answerable using health impact evaluations; for example, 'interventions to improve water quality for preventing diarrhoea' (Clasen et al., 2015). In recent years, there has also been a movement towards reviews covering multiple research questions

[33] Sandy Cairncross, pers. comm.

answerable using different types of evidence, such as 'effectiveness and factors influencing implementation of handwashing and sanitation promotion' (de Buck et al., 2017). Broader reviews enable greater statistical precision and systematic analysis of bias, as noted by Gøtzsche (2000): "[a] broad meta-analysis increases power, reduces the risk of erroneous conclusions, and facilitates exploratory analyses which can generate hypotheses for future research" (p.586).

Figure 3.13 Number of WASH systematic reviews by publication year



Notes: dotted line shows the end of the International Year of Sanitation (2008). The apparent decline in production of studies post-2018 reflects the limited searches done in this map after 2018.

A related issue is whether to set the question around an outcome – for example, 'water, sanitation and hygiene to tackle childhood diarrhoea morbidity in low- and middle-income countries' (Fewtrell and Colford, 2004; Waddington et al., 2009; Cairncross et al., 2010; Wolf et al., 2014, 2018) – or an intervention – 'effect of handwashing on infectious diseases' (Aiello et al., 2008). Some would further delimit by combining the two; for example, 'effect of handwashing on diarrhoea' (Ejemot-Nwadiaro et al., 2015), or perhaps 'the effect of improved water supply on women's time use' (a review which remains to be undertaken). But others might argue that hygiene can have a broader range of benefits in fighting respiratory infections (Rabie and Curtis, 2006; Mbakaya et al., 2017), and so should not be assessed

on its impact on diarrhoea alone. This debate amongst reviewers is known as 'lumping' versus 'splitting' (Gotzsche, 2000). One area where there does appear agreement is on the splitting of evidence collected in endemic versus epidemic conditions, since the effects of WASH in disease outbreaks are known to be much larger (e.g., Curtis and Cairncross, 2003; Gundry et al., 2004). This also includes WASH in emergency situations, where separate reviews have been completed (Brown et al., 2012; Yates et al., 2017).[34]

There is a tradition of measurement of intermediate and health outcomes in WASH impact evaluation, hence reviews have collected outcomes at different points along the causal pathway, examining contamination of drinking water between source and point-of-use (Wright et al., 2004), adherence to drinking water treatment and reported disease (Arnold and Colford, 2007) and differences in outcomes due to behaviour change (Waddington et al., 2009). 'Second generation' systematic reviews of interventions aiming to alter behaviour and measure broader behavioural and socioeconomic outcomes, are starting to appear. These include reviews of interventions like privatisation (Devkar et al., 2013). Some draw on broader evidence than impact evaluations, including process evaluations and qualitative studies, to understand factors determining implementation fidelity and reasons underlying adherence by participants (de Buck et al., 2017; Venkataraman, 2018). A few reviews include behavioural and socioeconomic outcomes. For example, Waddington et al. (2009) reported on diarrhoea studies that measured time-use, although did not specifically search for them, Annamalai et al. (2016) searched for evaluations of time use and Null et al. (2012) focused on willingness-to-pay.

Updates of reviews are becoming common as the evidence base expands. Systematic review updates have been done for Cochrane of household water treatment (Clasen et al., 2015) and hand hygiene (Ejemot-Nwadiaro et al., 2015). The review on WASH and diarrhoea infection (Esrey et al., 1985) has now been updated at least five times (Esrey et al., 1991; Fewtrell et al., 2005; Waddington et al., 2009; Cairncross et al., 2010; Wolf et al., 2014; Wolf et al., 2018). A criterion for updating a review is to update the searches for studies published more recently. But updates can usefully update other areas of a review, such as its scope (e.g., additional outcomes or sub-groups),

[34] A separate Cochrane group, Evidence Aid, exists to coordinate humanitarian evidence.

quality (e.g., methodological improvements, such as more comprehensive risk-of-bias assessment) and engagement (e.g., more comprehensive stakeholder consultation) (Waddington et al., 2018). For example, reviews of health impacts are incorporating analysis of participant adherence (Clasen et al., 2015). High quality synthesis of studies from existing impact evaluations, such as community-driven approaches, microfinance, and WASH in schools, as well as time-savings associated with water and sanitation improvements are needed. A systematic review update is urgently needed of the effects of water supply and hygiene on respiratory infections. Finally, a major omission from the current systematic review evidence base is the lack of a review focusing on the impacts of WASH interventions on mortality, whether all-cause or cause-specific, such as due to diarrhoeal disease. This synthesis gap is addressed in Chapter 6.

3.5 Ethics in WASH impact research

This section examines three ethical questions associated with the studies included in the WASH evidence census: rigour, or the quality with which studies are designed and implemented; relevance, the extent to which they answer important questions; and representation, how inclusively they have been conducted.

3.5.1 *Rigour*

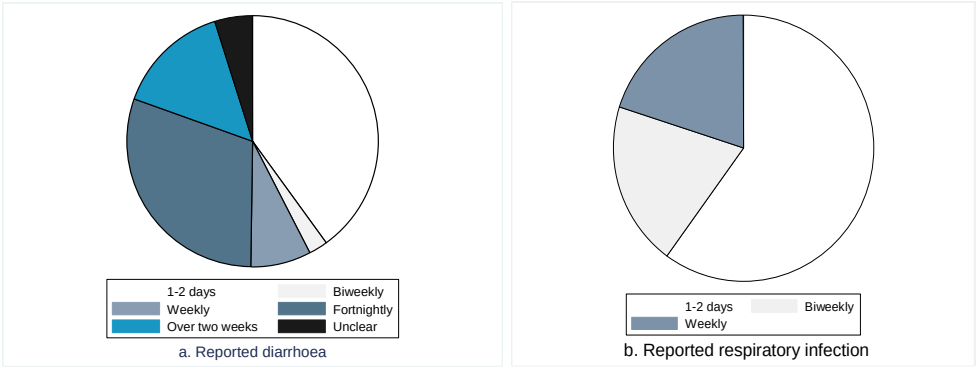
Mark and Lenz-Watson (2011) view research quality through an ethical lens, arguing that the wrong answer may result in harm to subsequent programme participants, where getting the wrong answer (or answering the wrong question) is largely due to limitations in study design and implementation. It is therefore important to get the right answer to the right questions, using the best available methods. There have been concerns about the quality of WASH impact evaluation at least since Blum and Feachem (1983) presented six areas where diarrhoeal health impact evaluation designs were suboptimal: use of a control group, adjustment for confounding, definition of the outcome, length of recall, analysis of use, and sample size. The impact evaluation evidence census suggests these points have been incorporated into common practice by WASH researchers. Thus, all studies used control or comparison groups that received no, or a different, intervention, with the exceptions of Duflo et al. (2015) who used interrupted time series to measure

infectious diseases following household water connections, and Arku (2010) who measured time use by participant recall before and after installation of improved community water supply. As noted in Chapter 1 Section 1.5, before-versus-after design is the preferred approach to measuring immediate outcomes like time savings from WASH improvements where there is no risk of confounding (Victora et al., 2004).

Almost all studies addressed confounding, either through random assignment, group or individual level matching on observables prior to analysis, or directly in adjusted analysis. For example, most studies now use centrally administered randomisation, although there is the occasional exception where a study has used quasi-randomisation through alternation (Montgomery et al., 2016). Some studies used randomisation over small samples, such as Stone and Ndagijimana (2018) who randomised across two districts in Rwanda. In non-randomised studies using matching, the matching was usually done using statistical methods, although a few used 'naïve' matching (e.g., World Bank, 1998). However, very few non-randomised studies have used rigorous methods to address unobservable confounding, such as double differences, interrupted time-series and regression discontinuity.

Outcomes were nearly always clearly defined for diarrhoea (95% of cases) usually being the WHO definition of "three or more loose stools in a 24-hour period", and where the diarrhoea incidence was reported "three intervening diarrhoea-free days" were required to define a new episode (Bacqui et al., 1991). For self-reported diarrhoeal disease, only a minority of studies used recall periods longer than two weeks (Elbers et al., 2012; Galiani et al., 2009; Iijima et al., 2001; Pradhan et al., 2002; Walker, 1999). Studies measuring respiratory infection by self-report used recall periods of, at most, seven days (Figure 3.14).

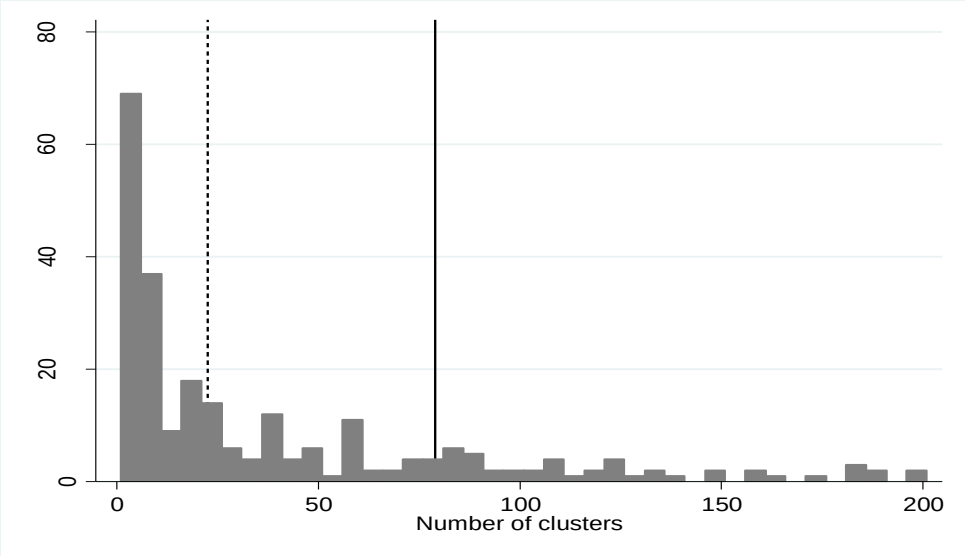
Figure 3.14 Recall period for self- or carer-reported disease



It is necessary to go beyond 'bare bones' by collecting data to answer relevant questions about implementation and causal mechanisms, not just on effects (Mark and Lenz-Watson, 2011). Use of causal pathway analysis is well-established and was done from the earliest trials of hygiene (e.g., Torún, 1982) and water treatment technology (e.g., Kirchhoff et al., 1985). Over half of studies collected data on behavioural outcomes. However, reporting of the WASH technology and intervention components (e.g., whether hygiene promotion was a component, frequency of contact between promoter and participant) was not always clear (see also Pickering et al., 2019).

Study sample sizes have also increased with the greater research resource availability. The median number of clusters across the sample is 21 (and the mean 79), whether cluster is defined as communities, villages, informal settlements, neighbourhoods, municipalities, schools or health facilities (Figure 3.15). For example, until 2008 the median number of clusters was only 10 (the mean was 49), whereas post-2008 it was 31 (mean of 92). Less than a quarter of studies published since 2008 have 'one-to-one' comparison (Blum and Feachem, 1983) effective sample sizes of less than ten clusters. More studies are therefore able to estimate statistically precise effects, over bigger samples which can provide useful information about scale and scalability, all of which are vital for policy relevance.

Figure 3.15 Frequency of WASH studies by cluster sample size



Note: dashed line shows the median, solid line shows the mean.

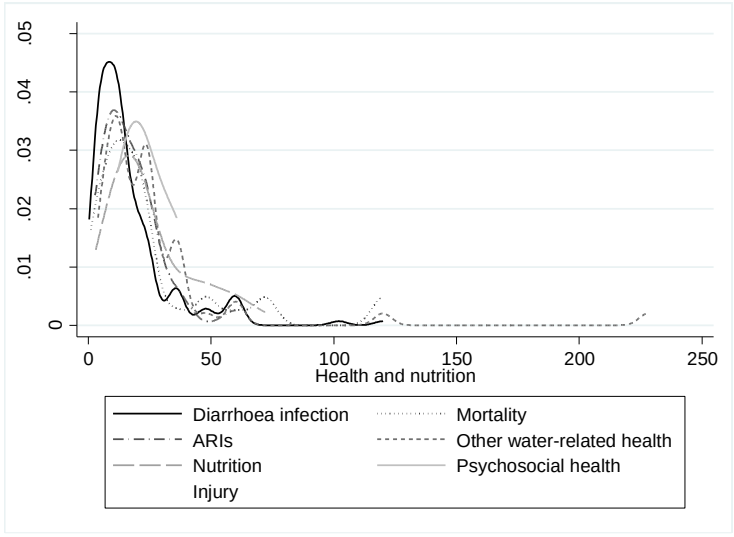
Sustainability, measured as sustained behaviours or quality of life outcomes, is also important for policy (e.g., Waddington et al., 2009). Data were collected on follow-up length, measured as the number of months from baseline or intervention inception to final follow-up, which varies by intervention (Table 3.3) and outcome (Figure 3.16). Studies of direct provision and health education, or those measuring diarrhoea and acute respiratory health outcomes, or water treatment and hygiene behaviours, were conducted over relatively shorter periods, with a median number of 12 months each. In contrast, studies of supply-side interventions such as decentralisation (e.g., community-driven development, median 24 months) or those measuring socioeconomic outcomes, which may take longer to materialise as they are further down the causal pathway than behaviours and health, tend to be conducted of longer follow-ups (median of 19 months for education outcomes, 30 months for income, and 48 months for labour market outcomes). Researchers and funders appear to have been sensitive to calls for greater examination of sustainability of interventions and outcomes (e.g., Waddington et al., 2009). For example, evaluations of CLTS, all of which were published since 2012, include studies measuring open defaecation several years after implementation – four years in the case of Adank et al. (2017), and ten years for Orgill (2017), which also measured education outcomes. The increased value in longer follow-up periods is well-recognised as a necessary check on slippage (Adank et al., 2017).

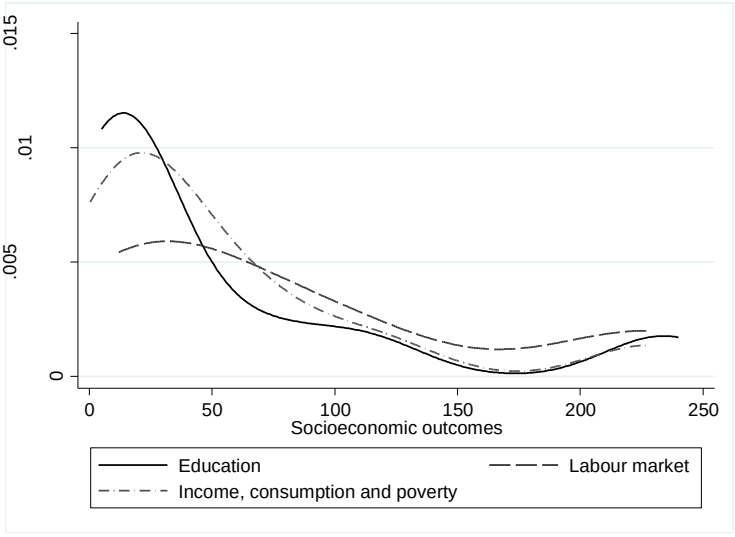
Table 3.3 Average length of follow-up (months) by intervention

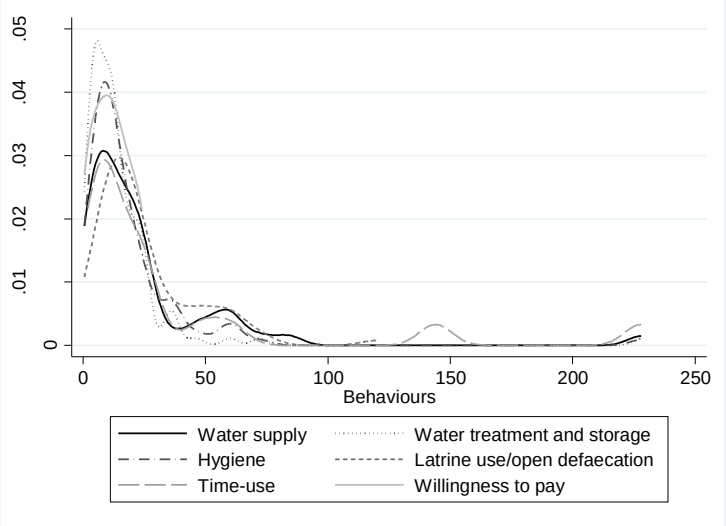
	Intervention	*Median*	*IQR*		*N*
Demand-side	Health education	12	6	24	130
	CLTS	24	12	36	24
	Other psychosocial triggering	12	8	18	45
	Subsidy	12	6	21	33
	Microfinance	22	18	24	6
	Legal reform	60	60	60	1
Supply-side	Direct provision	12	6	20	182
	Privatisation	84	30	180	4
	Small-scale independent provider	24	12	36	13
	Operator performance	21	18	24	0
Demand- and supply-side	Decentralisation (e.g., CDD)	24	12	54	23

Notes: IQR inter-quartile range; N number of study arms with any intervention component.

Figure 3.16 Months of follow-up by outcome (densities)

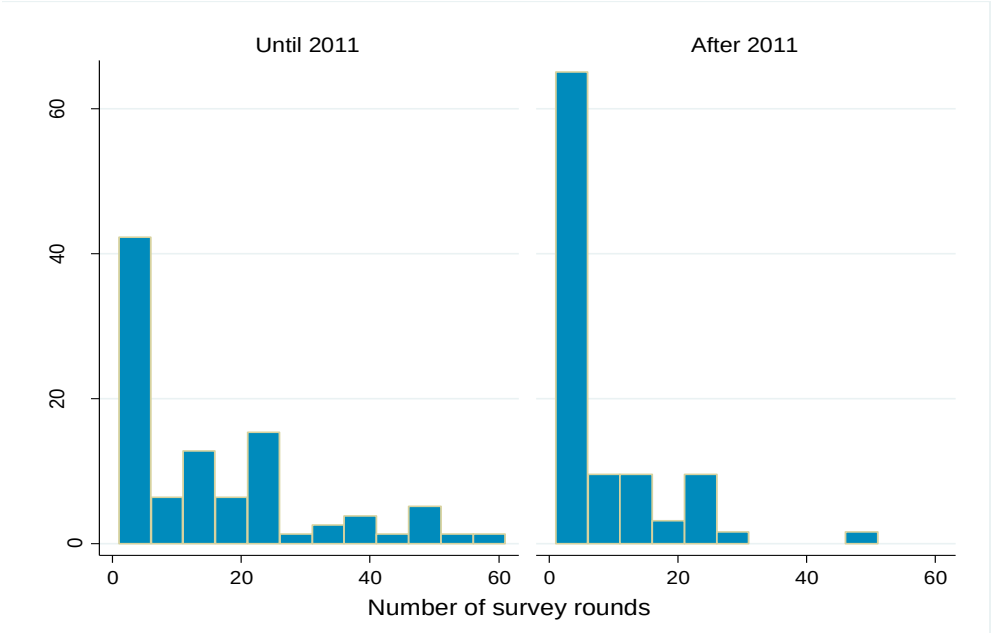






As a final measure of quality, data were collected on the number of survey rounds for health impact studies measuring self-reported diarrhoea (Figure 3.17). The average number of rounds of outcomes data collection has also fallen since the publication of papers suggesting significant bias in repeated measurement due to participant fatigue (Zwane et al., 2011).

Figure 3.17 Number of survey rounds in diarrhoea studies (%)



3.5.2 *Relevance*

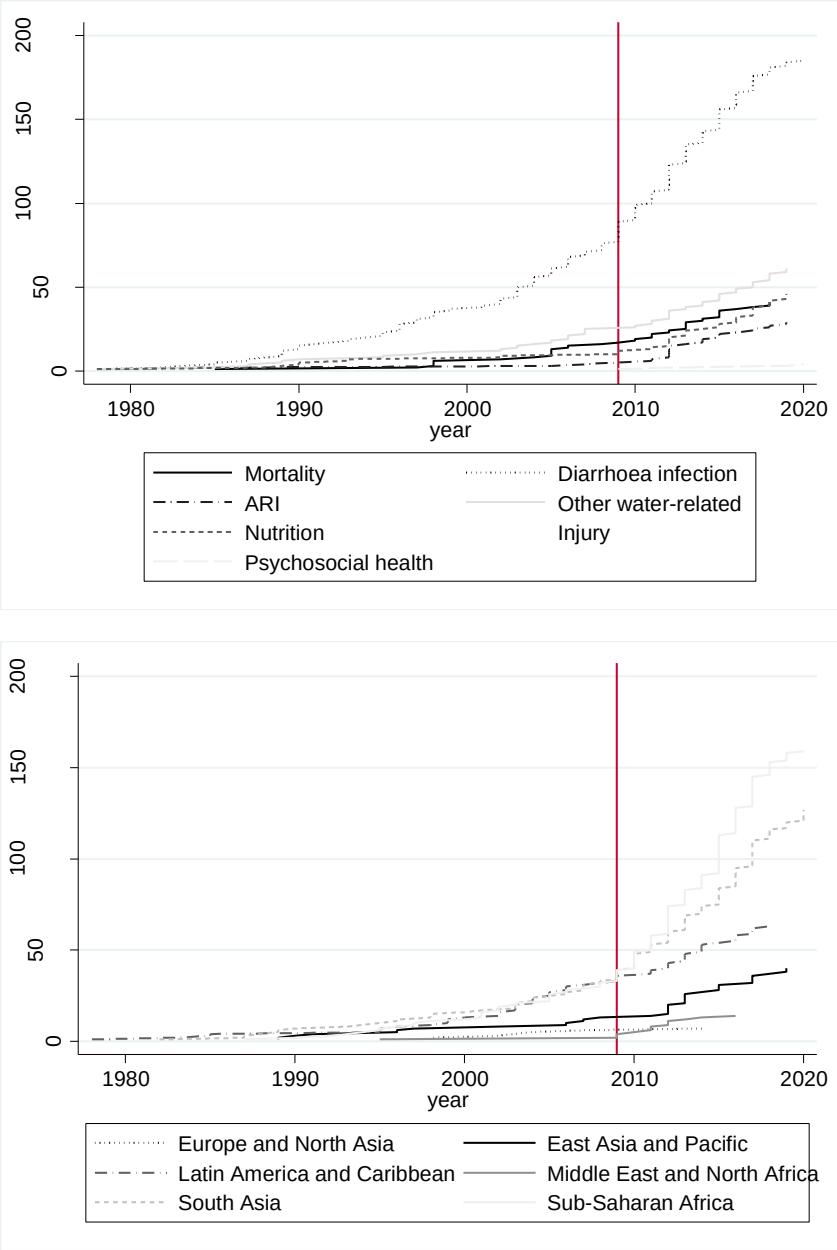
The most recent global burden of disease (GBD) exercise estimated 1.6 million deaths and 105 million DALYs were attributable to inadequate WASH annually (Prüss-Ustün et al., 2019). Of these, over 800,000 deaths

and 50 million DALYs were caused by diarrhoea, half of which were in sub-Saharan Africa and around one-quarter in South and East Asia. These are likely underestimates, as the figures on GBD attributable to WASH omit non-communicable diseases (e.g., arsenicosis or musculoskeletal disease) or sources of DALYs like injury, drowning, neonatal conditions and maternal outcomes. While estimates do not appear to have been produced to attribute these sources to WASH conditions, these are undoubtedly significant sources of global DALYs.[35] For example, 82 million DALYs were caused by road injury, 50 million were due to back and neck pain, 40 million due to neonatal sepsis and infections, and 20 million by drowning (as compared to 130 million due to acute lower respiratory infection and 81 million due to diarrhoea) (WHO, 2018).

An instructive comparison can be made of the distribution of WASH studies in L&MICs by outcome and location (Figure 3.18), according to the priorities given by the GBD. Table 3.5 presents data on the relationship between total sample size, disability-adjusted life years (DALYs), years of life lost (YLL) and years living with disability (YLD). Analysis suggests a positive correlation of total sample size with DALYs overall (Pearson rho=0.37), and for YLLs (rho=0.41), but a negative correlation with YLD (rho=-0.13), shown graphically in Figure 3.19. The latter is due to the limited number of studies measuring impacts on musculoskeletal disorders and psychosocial health.

[35] For example, Prüss-Ustün et al. (2008) estimated 280,000 preventable deaths annually due to drowning.

Figure 3.18 Cumulative total number of studies



Note: vertical line marks the end of the International Year of Sanitation (2008).

Table 3.4 WASH impact evaluation sample size and GBD estimates

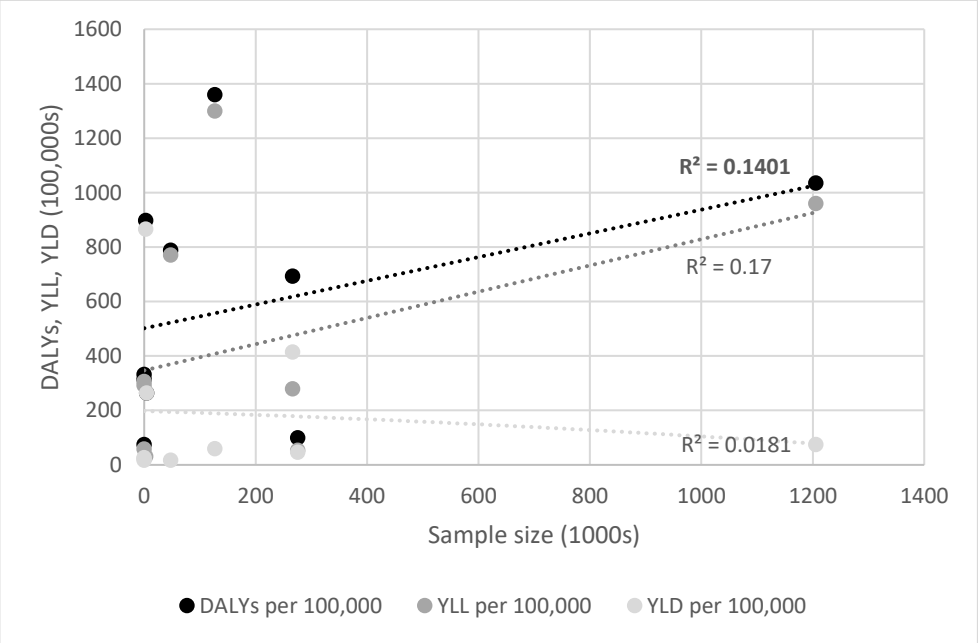
	Sample size	*DALYs*	*YLL*	*YLD*
Diarrhoea	1,205	1,035	960	75
Other water-related ill-health	275	99	53	46
Nutrition	267	693	279	414
Acute respiratory infection	126	1,359	1,300	59
Psychosocial health	5	264	-	264
Pedestrian transport injury	0	315	292	23
Musculoskeletal	3	897	31	866
Neonatal sepsis	0	332	306	26
Malaria	47	788	771	17

Animal contact	0	75	58	17
Pearson rho		0.37	0.41	-0.13

Note: sample size in 1,000s; DALYs, YLL and YLD in 100,000s. YLD due to psychosocial health attributed to anxiety. Other water-related ill-health indicators attributed to intestinal nematode infections and trachoma.

Source: data from GBD (2017a, 2017b).

Figure 3.19 Correlation between GBD and study participation



In addition, the correlations between the distribution of study participants and the regional distribution of GBD by outcome are strong for nutrition and, to a lesser extent, diarrhoea, but weak for other water-related ill-health (intestinal nematodes and trachoma) and respiratory infection (Table 3.5).[36] The correlations are generally stronger for RCTs than other studies, with the exception of respiratory infection where the correlation between numbers of participants and GBD by global region is very low (rho=0.10).

The economic benefits of WASH improvements, due to averted deaths, improved health, health care savings and time savings far exceed the costs of provision. For example, Hutton and Haller (2004) estimated the economic value of time savings to dwarf the estimated economic benefits due to diarrhoea, contributing to 65 percent of the benefits (as compared to around

[36] However, the correlations between regional GBD and number of study arms are weaker (rho=0.41 for the total GBD, rho=0.49 for diarrhoea, rho=0.32 for ARIs, rho=0.10 for other water-related illness), with the exception of nutrition (rho=0.84).

10 percent for days lost due to diarrhoea).[37] Later estimates confirmed that the majority of economic benefits from both water and sanitation were time savings (Hutton, 2015), although health benefits from improved water supply due to less diarrhoeal disease were revised upwards due to findings from a revised systematic review (Wolf et al., 2014).

Table 3.5 DALYs (per 100,000) by location and outcome

Total	*Diarrhoea*	*ARIs*	*Other water-related ill-health*	*Nutrition*	*Total*
Eastern Europe and North Asia	10	65	0.01	19	3,004
East Asia and Pacific	25	133	15	35	5,104
Latin America and the Caribbean	21	90	3	31	2,869
Middle East and North Africa	90	158	2	57	2,511
South Asia	248	339	13	262	7,125
Sub-Saharan Africa	422	606	11	254	5,986
Pearson rho (all studies)	0.49	0.32	0.10	0.84	0.42
Pearson rho (RCTs)	0.65	0.10	0.43	0.78	0.85

Notes: Other water-related ill-health attributed to intestinal nematode infections and trachoma. Pearson correlations with sample size by location and outcome.

However, the estimates for economic benefits of WASH provision are usually estimates of opinionated experts or minimum wage data (Hutton and Haller, 2004; Hutton, 2015) and occasionally observational studies in the case of time savings (Hutton et al., 2007). They are not based on observed benefits measured in impact evaluations (White and Gunnarson, 2008). Despite the clear economic value of improved WASH, and the strong negative correlation between total study sample size and benefits (rho=-0.31), only a small share of evaluations has been able to measure socioeconomic outcomes.

Another perspective comes from those at the bottom, the users of WASH services. For example, a survey of women in Benin (Jenkins, 1999) found that commonly perceived benefits of sanitation were safety and comfort (Table 3.6), whereas health was rarely mentioned. The Pearson correlation between

[37] Hutton et al. (2007) also estimated the global distribution of economic benefits for improved water and sanitation, 36 percent were in the Western Pacific region (including China), 24 percent in Latin America and the Caribbean, 19 percent in South and South-East Asia (including India), 9 percent in sub-Saharan Africa, and 4 percent in the Eastern Mediterranean and 4 percent in Central and Eastern Europe.

outcomes collected in L&MIC WASH research and average scores by participants in Jenkins (1999) is strongly negative (rho=-0.79). Clear opportunities should be taken to fill these research gaps.

Table 3.6 Reasons given for the benefits of sanitation in Benin

Reason	*Outcome construct*				*Score*
	Safety	*Status*	*Comfort*	*Health*	
Avoid discomforts of the bush	Y				3.98
Gain prestige from visitors		Y			3.96
Avoid dangers at night	Y				3.86
Avoid snakes	Y				3.85
Reduce flies in compound			Y		3.81
Avoid risk of smelling/seeing faeces in bush			Y		3.78
Protect my faeces from enemies	Y				3.71
Have more privacy to defecate			Y		3.67
Keep my house/property clean			Y		3.59
Feel safer	Y				3.56
Save time			Y		3.53
Make my house more comfortable			Y		3.50
Reduce my household's health care expenses				Y	3.32
Leave a legacy for my children		Y			3.16
Have more privacy for household affairs			Y		3.00
Make my life more modern		Y			2.97
Feel royal		Y			2.75
Make it easier to defecate due to age or sickness			Y		2.62
For health (spontaneous mention)				Y	1.27
Be able to increase my tenants' rent		Y			1.17
Average score	3.79	2.80	3.44	2.30	

Note: Y=reason relates to outcome construct.

Source: Jenkins (1999).

3.5.3 *Representation in WASH research and research governance*

Over thirty years ago, Cairncross (1989) stated that the fundamental aspect of WASH evaluation research was that it needed to be conducted in the low- and middle-income country environments where water, sanitation and hygiene programmes were implemented. He stated that "[t]his means that it should ideally be conducted by developing country nationals" (p.308) who, all else equal, have better knowledge of the contexts in which programmes are implemented, and also have better knowledge of, and ties to, those taking decisions about programming in-country (and possibly also programme participants). These studies should have a better chance of uptake by

decision-makers and therefore in improving lives. However, he noted, the international agencies that to their credit sponsor research into developing sanitation technologies or evaluating WASH programmes tended to employ Western experts, and "very little effort" (p.308) was made to develop research capacity in L&MICs.

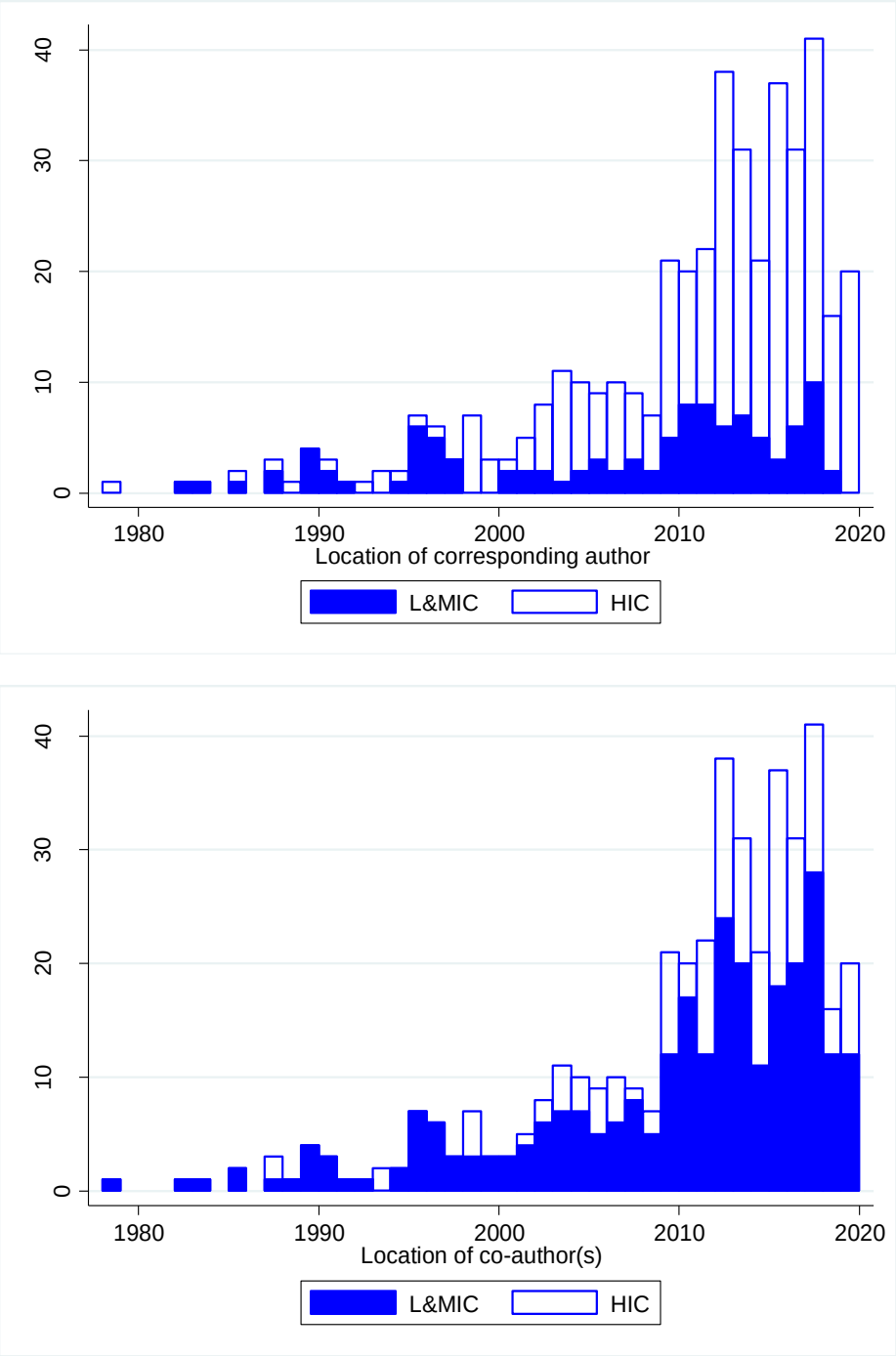
Some of the earliest rigorous WASH trials were led by L&MIC researchers, such as Khan's (1982) factorial study of handwashing and water treatment and storage in Bangladesh, the crossover trial of household water treatment by Kirchhoff et al. (1985) in Brazil, as well as RCTs of handwashing in Myanmar (Han and Hlaing, 1989), and a factorial trial of filtration and handwashing in Guatemala (URL, 1995). This suggests a high degree of representation of L&MIC authors in early impact evaluations. To what extent has this changed?

Data were collected on institutional location of lead or corresponding authors and co-authors of WASH impact evaluations. Figure 3.20 plots the evolution of all impact evaluations according to whether the lead or corresponding author or at least one co-author, were based at an institution in the L&MIC where the study was conducted, or in a high-income country (HIC). While research leadership in L&MICs has increased over the period, with the increased resources available for research in the sector as a whole, it has not increased as appreciably as a proportion of total studies. If anything, there has been a deterioration since the 1980s and 90s when the majority of WASH impact evaluations were led by L&MIC researchers.

Figure 3.22 plots the same data for RCTs. There has been a marked increase in co-authors based in L&MICs, to the extent that it is more common for authorship to include at least one L&MIC co-author than not. In most cases, however, this is a single L&MIC researcher on a paper with four or more co-authors, whose role does not appear to be one of study design, data analysis or writing up. Rarely, the corresponding author and most (e.g., Messou et al., 1997) or all co-authors (e.g., Garba et al., 2001; Roushdy et al., 2011; Ozcelik et al., 2013; Makotsi et al., 2016) are from an L&MIC institution. Another study found that rates of authorship from the country of investigation in clinical trials was much lower in L&MICs than in HICs, for example around 30 percent in Brazil and India and as low as 13 percent in Peru (Hoekman et al., 2012). Echoing these findings, a cross-sectoral scoping study recently

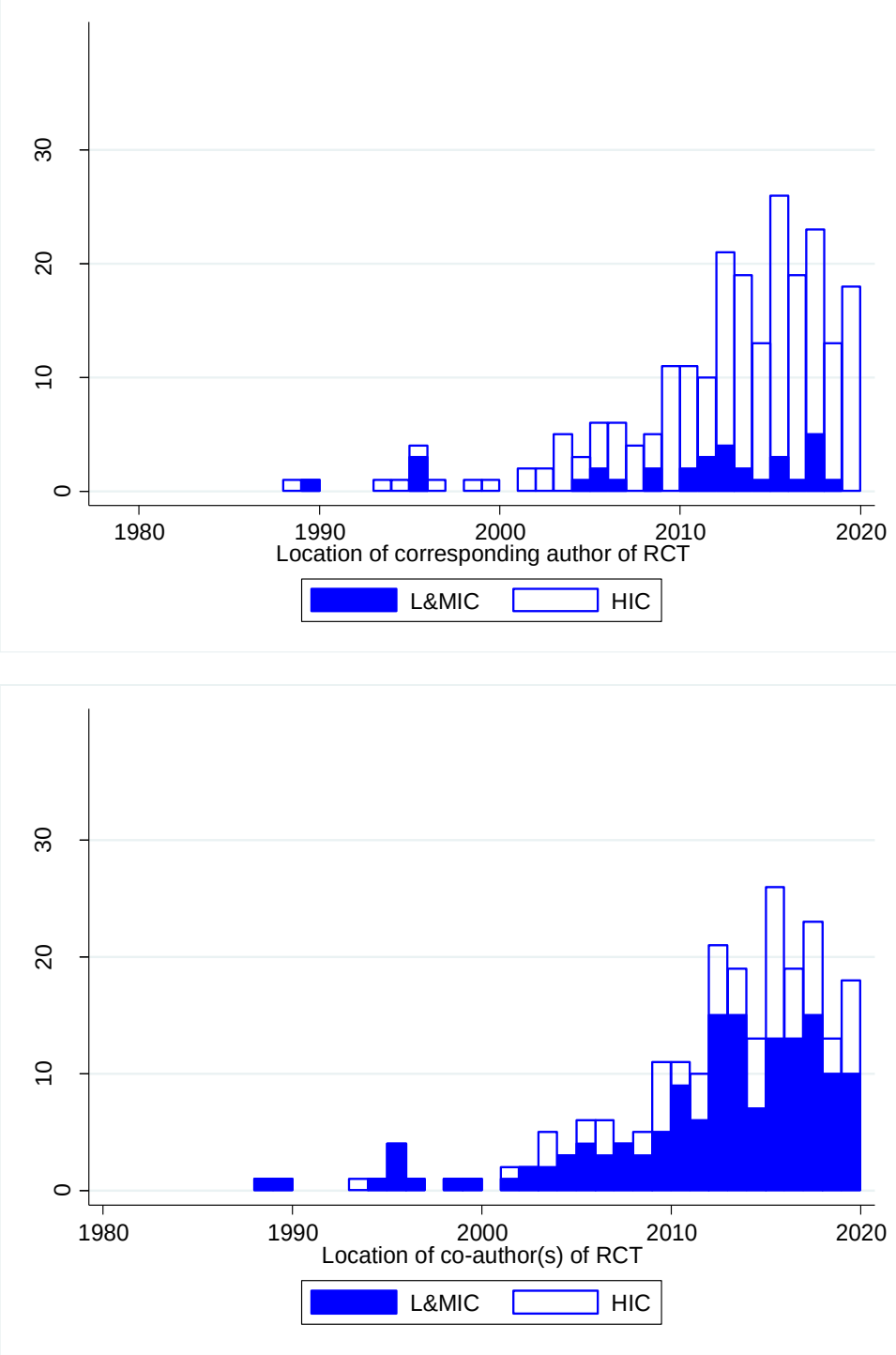
found 1,500 African researchers had been involved in impact evaluation publications between 1990 and 2015, but only 13 percent were first authors and in only 2 percent of studies were all authors based in African institutions (Erasmus and Jordaan, 2019).

Figure 3.20 Number of WASH studies by author location



Note: co-author(s) listed as L&MIC if at least one co-author is based there.

Figure 3.21 Number of WASH studies by author location – RCTs



Note: co-author(s) listed as L&MIC if at least one co-author is based there.

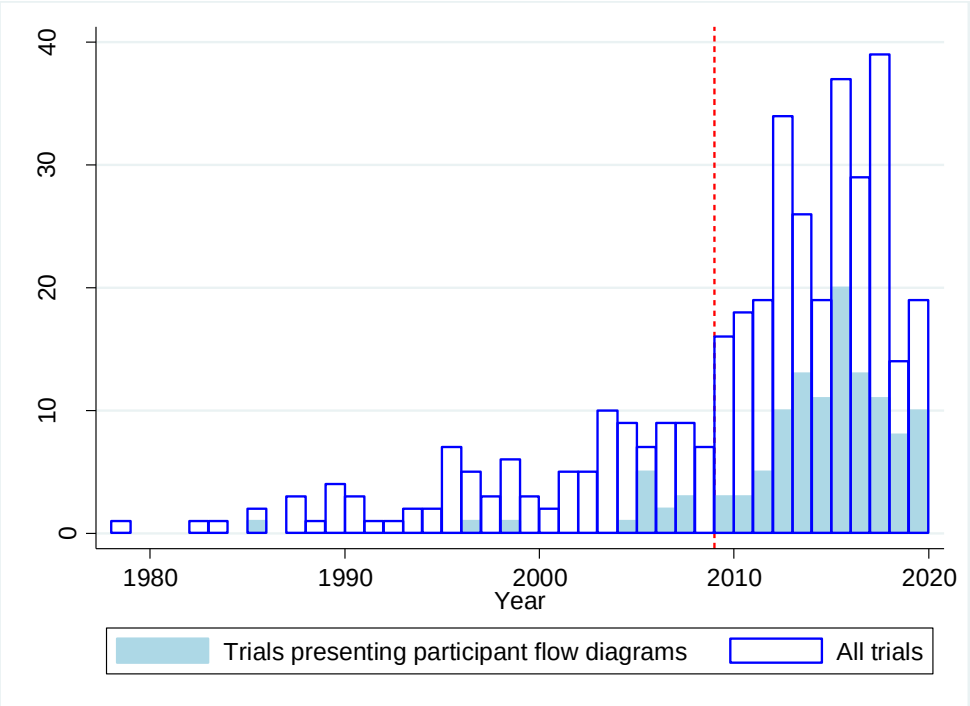
The situation in systematic reviews and evidence synthesis is also changing. The first reviews were done by Steve Esrey (1985, 1991), and later many were led by researchers in Western institutions. However, there have been some international efforts to institutionalise systematic reviewing in L&MICs since at least 2007, when the WHO Alliance for Health Policy and Systems Research centres were established in Bangladesh, Chile, China and

Uganda.[38] More recently, the Africa Evidence Network, coordinated by the Africa Centre for Evidence (ACE) at the University of Johannesburg, was set up with aim of promoting evidence-informed decision-making including through synthesis work.[39] The Global Evidence Synthesis Initiative (GESI), based at the American University in Beirut, was established to promote systematic review supply and demand in L&MICs; its network contains 47 evidence synthesis centres from 25 countries.[40] The Campbell Collaboration opened a South Asia office in New Delhi in 2015.[41] All are very welcome initiatives, but more could be done, especially now with technological improvements potentially available for remote working, if major funders – and possibly also journals[42] – were to incentivise it. However, some of the challenges remain fundamental. As noted in a Lancet editorial, "many of us [L&MIC researchers] are experiencing common difficulties arising from limited access to computer hardware and software, restrictions on database access, limited data storage capacity, inadequate data coverage, and low internet bandwidth" (Stewart et al., 2020, p.2).

There are also reasonable questions about research governance. As noted by White (2013), "[t]here has been an enormous increase in data collection in developing countries in the last decade. Surveys are time consuming for respondents. So, we have to really believe that what we are doing is worthwhile not just for us, but for the poor people whose time we are taking in conducting our studies. This consideration seems not to weigh heavily with many researchers, but clearly it should…" (p.47). Unfortunately, current standards for reporting, especially in social science (mainly development economics) working papers and journals, are poor. As shown in Figure 3.22, the basic requirements of reporting participant flow adherence in field trials according to CONSORT standards (Moher et al., 2010) have improved over time but are frequently unmet.

[38] https://www.who.int/alliance-hpsr/researchsynthesis/project2/en/ (accessed 9 October 2020).
[39] https://africacentreforevidence.org/ (accessed 9 October 2020).
[40] http://www.gesiinitiative.com/about-gesi (accessed 24 October 2020).
[41] https://campbellcollaboration.org/southasia/ (accessed 9 October 2020).
[42] For example, Tropical Medicine and International Health editors required papers to have at least one L&MIC co-author (Sandy Cairncross, pers. comm.).

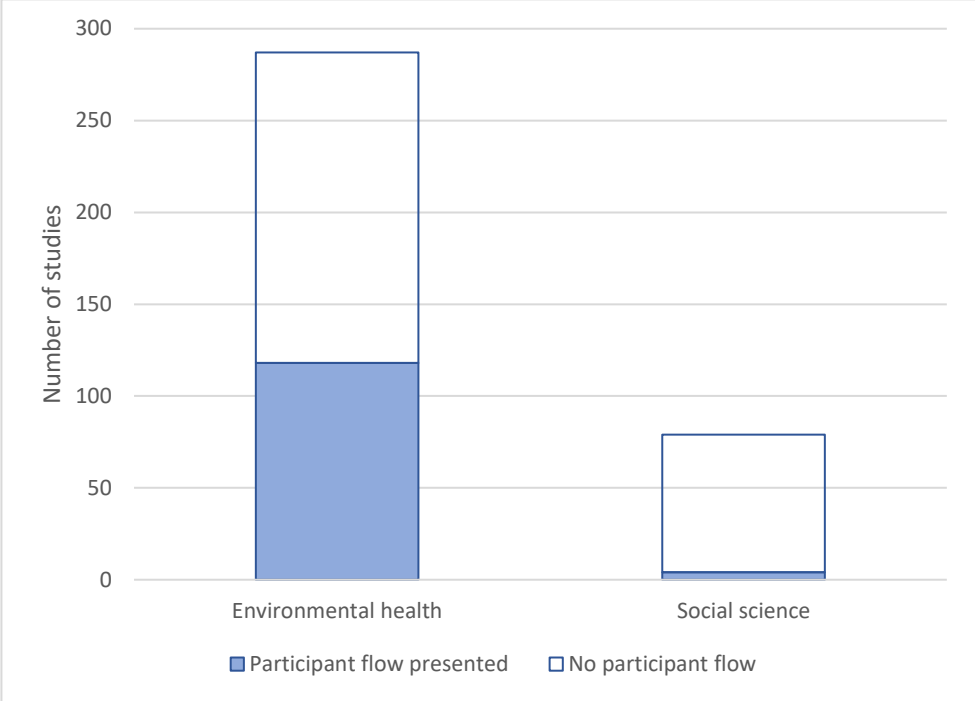
Figure 3.22 Number of trials presenting participant flows by year



Notes: dotted line marks the end of the International Year of Sanitation (2008). The apparent decline in production of studies post-2018 reflects the limited searches done in this map after 2018.

If the reporting in environmental health is substandard, with less than 50 percent of studies presenting participant flows, the reporting in social sciences may go as far as being deliberately misleading (Figure 3.23). Only two out of 54 prospective studies in social science presented a participant flow diagram or the data from which it could be fully reconstructed (Beath et al., 2013; Guiteras et al., 2015a). Partial exceptions were Kremer et al. (2008) and Okyere et al. (2017) – which both provided aggregated numbers of participants at follow-up, not by study arm – as well as Jalan and Somanathan (2008) and Malek et al. (2016). In addition, Orgill (2017) provided detailed analysis of household attrition by survey round and treatment group, from which participant flow could be determined. Others provided truncated flow diagrams, excluding participation flow data in follow-up periods (e.g., Doocy and Burnham, 2006; Jalan and Somanathan, 2008; Biswas et al., 2012; Malek et al., 2016).

Figure 3.23 Participant flow diagrams by academic discipline



Some newer studies by social scientists are starting to exhibit flow diagrams for the full trial period, at the cluster level, but not yet at individual level (e.g., Armand et al., 2020). This lack of transparency makes it difficult to appraise study validity, as well as inhibiting the use of important information that can be used in synthesis work for policy audiences, such as analysis of all-cause mortality as shown in Chapter 6. It is clear that these failures stifle scientific progress, and WASH triallists should accept as good practice standards adopted in clinical epidemiology decades ago (Moher, 1998).

Data were also collected on ethical review reported in WASH impact evaluations (Table 3.7). Again, while standards in environmental health, of which over half of studies that would need ethical review, could be improved, the standards in social science leave much to be desired. Only 22 percent transparently indicated an institutional review board (IRB) had approved the evaluation, and even fewer (16%) had done so at IRB in-country; nothing was indicated about ethical review in 67 percent of cases. In over 10 percent, no ethical review appeared to have been followed. Thus, no study published by a UN body, whether the World Bank, a regional development bank, UNICEF or other organisation indicated that an institutional review process was undertaken prior to study implementation.

Table 3.7 Ethical review in WASH impact evaluations (%)

	Total	*Environmental health*	*Social science*
Passed any IRB	43	55	22
o/w passed IRB in country	37	47	16
No IRB was consulted	6	5	11
Unclear/not stated	49	39	67

Note: may not sum to 100% due to rounding errors.

It is possible that programme evaluations, which are the studies conducted by development banks, are thought not to require ethical approval, as they are being rolled out anyway. For example, Semenza et al. (1998) indicated that "IRB review was not required because the study did not fall under the human subjects regulations" (p.941) as it was a programme evaluation. This was despite the evaluation including a component where participants were randomised to receive chlorine and a safe storage device. In this case, and in the cases of prospective evaluations done by the development banks, there may be ethical issues relating to withholding treatment from control communities, or the ethical standards around, for example, compensating participants for their time, and possibly by offering health treatment to the severely ill, such as oral rehydration salts for diarrhoea.

3.6 Conclusion

To summarise, there has been a dramatic increase in quantity and focus of impact evaluations and systematic reviews on WASH topics. There has been a movement to broaden the range of outcomes beyond diarrhoeal disease in WASH impact evaluations and systematic reviews, corresponding to a 'behavioural revolution'. Other health and socioeconomic outcomes are likely to be more important in determining acceptability, and therefore household demand for, new WASH technologies. For example, safety, status and convenience are all considered more important than health in determining sanitation demand. This chapter found that rigour in the conduct of evaluations and reviews has improved since the first International Drinking Water Supply and Sanitation Decade during the 1980s. It will be important that these standards are maintained through the second UN International Water Decade (2018-2028), to ensure resources for WASH programming are spent in the most effective way to achieve universal coverage. However, there are concerns about how relevant the studies are for

top-down or bottom-up sector priorities, and the clock appears to have stalled or even rolled back on representation of L&MIC study leadership, and there are important issues relating to ethical standards and reporting WASH sector impact research. Systematic reviews are often restricted to literature published in academic journals, a practice which would tend to bias the estimated impacts of WASH programmes, reducing confidence in findings.

It is striking how few studies have taken advantage of natural experiments to answer questions that prospective approaches like RCTs cannot, compared to other sectors (Dunning, 2012). Natural experiments, applying statistical methods of correction for unobservable confounding to existing surveys, remain an underutilised methodological approach in WASH evaluation. The large numbers of existing household survey datasets available containing questions on WASH exposures that are already being examined (e.g., Fink et al., 2011; Geere and Hunter, 2020) suggest great promise for these approaches. There also continues to be a great number of uncontrolled studies that simply measure outcomes before and after the intervention. Most of these studies were excluded as they are not usually able to attribute changes to the intervention, the exception being for the immediate outcomes of time savings due to provision of a new water supply or sanitation source, for which evidence synthesis is ongoing (Macura et al., 2021).

Chapter 4 A tool to assess fragility of inference in impact evaluation

> "The haphazard way we individually and collectively study the fragility of inferences leaves most of us unconvinced that any inference is believable. If we are to make effective use of our scarce data resources, it is therefore important we study fragility in a much more systematic way."
>
> Leamer (1983, p.43).

4.1 Introduction

Those producing WASH impact evaluation are primarily epidemiologists and social scientists, who quantify treatment effects – that is, measured changes in outcomes among populations exposed to an intervention, as compared to populations not exposed – using randomised and non-randomised study designs. Non-randomised studies include designs like regression discontinuity, interrupted time-series, non-equivalent comparison group designs like case-control, and methods of estimation like difference-in-difference, instrumental variables and multiple regression. They are also referred to variously as quasi-experiments (e.g., Shadish et al., 2002; Waddington et al., 2009; Bärnighausen et al., 2017a; Reeves et al., 2017), natural experiments (Craig et al., 2011; Dunning, 2012), or observational studies (e.g., Cook and Steiner, 2010).[43]

All quantitative causal studies are subject to biases relating to attribution (internal validity) and the extent to which findings are generalisable to the population and variables of interest (external validity) (Shadish et al., 2002). RCTs, often considered the preferred method of causal inference where they are feasible (e.g., Rubin, 1974; Shadish et al., 2002; Duflo et al., 2006; Imbens and Wooldridge, 2009), can have methodological problems in design and implementation such as poor allocation concealment, non-random attrition, contamination of controls, biases in analysis and reporting, and so

[43] Some authors have chosen not to highlight the differences. For example, Cook and Steiner (2010, p.57) stated that they use the terms 'quasi-experiments' and 'observational studies' interchangeably.

on (Higgins et al., 2011). Threats to internal validity due to participant knowledge about investigation, are thought to be more problematic in trials (whether randomised or otherwise) than observational studies, due to the process of informed consent (Schmidt, 2014). Threats to external validity are also thought of as being more problematic in trials due to modifications to usual treatment practice and/or closer monitoring of implementation (Bärnighausen et al., 2017b). Another issue with external validity in trials and some quasi-experiments[44] is that participants and interventions are usually chosen through convenience, rather than random sampling as they might be in a purely observational study based on a representative household survey (e.g., Pritchett and Sandefur, 2013).

Similarly, while non-randomised studies can produce the same effects as RCTs in meta-analysis (Concato et al., 2000), studies that are inappropriately designed or executed will not generate good causal evidence (e.g., Sacks et al., 1982). However, the threats to internal validity are often seen as more problematic, due to the greater risks of confounding, selection bias, and biases in analysis and reporting (e.g., Higgins et al., 2011; Sterne et al., 2016). The assessment of NRS design and implementation is also more difficult than RCTs, and tools are less advanced, requiring greater qualitative appraisal of potential biases, which in many cases may need to draw on advanced theoretical and statistical knowledge. Some types of observational studies popular among econometricians, so-called 'natural experiments', are viewed with particular suspicion. For example, referring to a recent natural experiment on the impacts of latrine provision on child diarrhoea mortality, Schmidt (2014, p.524) stated "India is colourful, but that is nothing compared to econometric analysis…". It is understandable that studies which purport to provide the 'holy grail' in solving the combined problems of bias in observational studies (due to confounding) and bias in trials (due to expectations effects) should be carefully assessed.[45] One may argue that part of the reason why natural experiments are viewed with suspicion is the lack of systematic critical appraisal which would enable others to assess the veracity of claims made in these studies.

[44] This includes studies producing any type of 'local average treatment effect' in which the estimate is valid for a subset of the population, such as those at the margin of the treatment threshold (in the case of regression discontinuity design) or compliers (in the case of instrumental variables estimation).
[45] Sampling bias is only really addressable when comparing findings across a large number of studies, or by using imputation methods to assess the likely effect in a particular context (e.g., Tipton, 2013).

These points are well understood in the policy research community. For example, the International Initiative for Impact Evaluation (3ie) Principles for Impact Evaluation states: "evaluation designs must be capable of addressing: a) confounding factors; b) selection bias; c) spillover effects; d) contamination of control groups; and e) impact heterogeneity by intervention, beneficiary type and context" (3ie, undated, p.2).

Systematic critical appraisal is therefore a key component of evidence synthesis work. Thesis Question 2 asks how to assess bias transparently and consistently for RCTs and NRS. Appraisal of internal validity, operationalised through 'risk of bias' assessment, gives assurance of the credibility of the point estimates provided in causal studies for the populations on which they are based (Higgins and Green, 2011) and, when combined with assessment of external validity, their credibility for the broader population and relevance for decision-making (Chalmers, 2014). Risk-of-bias tools aim to provide transparency about the judgments made by reviewers when performing assessments. They are usually organised around particular domains of bias and provide specific 'signalling questions' which enable reviewers to evaluate the likelihood of bias. Some tools are also operationalised to enable comprehensive validity assessment (Valentine and Cooper, 2008). Existing approaches, however, to differing degrees, are likely to provide misleading risk-of-bias assessments for randomised and non-randomised studies with selection on unobservables (Waddington et al., 2017). Nor is it clear whether they are developed or tested based on systematic evidence about bias (Villar and Waddington, 2019).

This chapter addresses Thesis Question 2 by discussing threats to validity in impact evaluations and operationalising a comprehensive risk-of-bias tool for randomised and non-randomised studies using statistical methods to identify causal relationships. Section 4.2 defines bias in relation to internal and external validity. Section 4.3 discusses ways of categorising impact evaluation, focusing on studies of WASH interventions. Section 4.4 discusses internal validity and Section 4.5 external validity. Section 4.6 presents proposed evaluation criteria for a critical appraisal tool to evaluate internal and external validity in randomised and non-randomised impact evaluations.

4.2 Conceptualising bias in impact evaluation

This chapter is primarily about three main threats to validity – how the observed effect may differ from the 'true' effect – in a study's findings: internal validity – that is, whether there is bias in estimating the 'true' effect for the sample; external validity – whether there is error in estimating the 'true' population effect, sometimes called sampling bias; and sampling error, measured as the standard deviation in the study estimate.

More formally, bias for study i is equal to the difference between the estimated effect – the sample mean $\hat{b}_i$, in impact evaluation called the average treatment effect (ATE) – and the 'true' target parameter – the population mean β, or population average treatment effect (PATE) (e.g., Greenland, 2000; Tipton, 2013):

$$bias_i = \hat{b}_i - \beta \qquad (4.1)$$

Bias is usually thought of as being determined by the study design and methods of implementation (for example, if the participants self-select to treatment and comparison, or if the measurement of outcomes is done inaccurately). However, the second component of bias, sampling bias, is determined by the way in which the study participants themselves are sampled (for example, whether participants themselves are randomly sampled from the population, whether the intervention being evaluated is chosen randomly, or whether sampling of either is done based on convenience). Hence, ATE and PATE are equal in expectation for an unbiased estimator, or equivalently the difference between them is zero, when a sufficiently large sample is chosen randomly from the target population. When the study draws on participants who are not randomly sampled from the population (e.g., participants or interventions are chosen for study due to convenience), as is standard in field research, ATE may be systematically different from PATE (sampling bias), although it still may provide an unbiased estimate of the sample ATE.[46] It is worth noting that an advantage of observational studies based on representative household surveys, over randomised field trials (and non-randomised treatment effect estimators) as usually implemented, is the reduced risk of sampling bias

[46] In an RCT where participants are selected based on convenience, the sample ATE may therefore be considered a population 'local average treatment effect' (LATE).

(Pritchett and Sandefur, 2013). In addition, Behrman and Todd (1999) refer to 'randomisation bias' (Heckman and Smith, 1995) where the process of randomisation generates changes in programme targeting – e.g., by lowering programme admission standards to meet sample size requirements – or population mobility – in the case of large-scale cluster controlled trials, where participants may be unwilling to migrate out of treatment clusters for fear of losing benefits[47] – which may make the findings inapplicable to the non-experimental context (see also Bracht and Glass, 1968).

The third property, the standard deviation of the estimator s_i measures the expected spread of mean values of the estimator from repeated random samples drawn from the target population, and largely depends on the study sample size:

$$s_i = \frac{\sigma}{\sqrt{n_i}} \qquad (4.2)$$

where σ is the sample standard deviation (that is, the sample-based estimate of the population standard deviation) and n_i the sample size for study *i*. There is therefore variance in an unbiased estimator in expectation, even if the random draws are from the same population, due to sampling error (sampling variation). This is usefully represented in two measures, statistical confidence and power. The confidence in the estimator – usually measured by the 95 percent confidence interval, associated with statistical significance level of $\alpha = 100 - 95 = 5$ percent – indicates that the 'true' effect is expected to lie within the interval in 95 out of 100 randomly drawn samples from the population:

$$\hat{b} \pm 1.96\, s_i \qquad (4.3)$$

where 1.96 is the critical value of the Z-distribution associated with $\alpha/2 = 5$ percent significance. Alternatively, there is an $\alpha = 5$ percent chance that the estimator will generate a false positive, wrongly concluding there is an effect when in fact there is not (also called Type I error). Another source of error occurs when the estimator wrongly concludes that there is no effect, when in fact there is (called Type II error). This is usually set at $\beta = 20$ percent,

[47] This is different from crossovers due to contamination, where control group units choose to migrate to treated communities to obtain benefits, which is a threat to internal validity.

indicating that there is a 20 percent chance of a false negative. Statistical power is the chance of correctly identifying a true positive, equal to $1 - \beta =$ 80 percent in the standard case.

Greenland (2000) states that "[e]stimators with large standard deviations (random scatter) are unreliable estimators of the target parameter, even if they are unbiased" (p.159). Hence, to get a fuller picture of the reliability of the estimator, one needs a measure incorporating both bias and standard deviation. One such statistic, measuring the expected average distance between the sample mean produced by estimator *i* and the population mean, is the mean squared error (MSE):[48]

$$MSE_i = bias_i^2 + s_i^2 \qquad (4.4)$$

where s_i^2 is the sampling error variance for estimate $\hat{b}_i$ (also called the variance of the effect), equal to square of the standard deviation:

$$s_i^2 = \frac{\sigma^2}{n_i} \qquad (4.5)$$

As discussed below, it is not clear what the effect of bias will be on the direction of bias. For example, while measurement error in independent variable (treatment) causes downwards bias in expectation (e.g., Wooldridge, 2009), measurement error in dependent variable (outcome) may upwards or downwards bias the estimate (e.g., courtesy or discourtesy bias in self-reporting), confounding may cause upwards or downwards bias depending on the relationships between omitted variable and dependent and independent variables, and so on.

However, where the samples come from heterogeneous sub-populations – for example, repeated replication studies based on samples drawn from populations with different characteristics – additional variation is expected over and above sampling variation, arising from differences in the treatment (e.g., intensity or length of administration), differences in outcome measurement (e.g., reliability in measurement), or differences in settings

[48] Since MSE is based on the squared deviations, it is sensitive to outliers. Other measures of average distance that are less sensitive include the mean absolute deviation and measures based on the median.

and potential outcomes for participants themselves (e.g., due to different demographic characteristics, such as age or sex, time period or season of data collection, or in the case of communicable disease, underlying environmental health risk). In theory, this may also include convenience samples, therefore accounting for sampling bias. All these factors cause variance in the 'true' population effect τ^2 (which is unobserved), over and above bias and within-study sampling error. In the case of heterogeneous sub-populations, therefore, the mean-squared error may be defined as:

$$MSE_i = bias_i^2 + s_i^2 + \tau^2 \qquad (4.6)$$

Because of these issues relating to bias and sampling error, it is usually agreed that lessons from policy research should be made using systematic methods of synthesis such as meta-analysis that "form a powerful, scientific approach to analyzing previous studies" (Littell et al., 2008, p.1). Meta-analysis, which is the statistical pooling of findings across studies, gives an estimate of the population parameter, by calculating an average effect across the estimates from single studies. By increasing the sample size, meta-analysis reduces the variation, increases precision and lowers the chances of Type I and Type II errors. Fixed effect meta-analysis calculates a pooled effect $\hat{\beta}_{FE}$ as the geometric mean where each effect is weighted by the inverse of its variance $\frac{1}{\sigma^2/n_i} = \frac{n_i}{\sigma^2} = w_i$. Since the weight for a single study is equal to the inverse of the variance, it follows that the variance of the fixed effect average s_{FE}^2 is the inverse of the sum of the weights across *k* included studies (Borenstein et al., 2009):

$$s_{FE}^2 = \frac{1}{\sum_i^k w_i} = \frac{1}{\sum_i^k \frac{n_i}{\sigma^2}} \qquad (4.7)$$

Fixed effect meta-analysis assumes that the studies are sampled from the same underlying population, with a single population average (PATE) and variance. Under the simplifying assumption of equal sample sizes, (4.7) can be rearranged as (Borenstein et al., 2009):

$$s_{FE}^2 = \frac{\sigma^2}{kn} \qquad if\ n_i = n_k = n \qquad (4.8)$$

The 95 percent confidence interval associated with the meta-analysis effect, represents the 95 percent likelihood that it incorporates the 'true' population parameter (equation 4.3).

Random effects meta-analysis, in contrast, assumes the studies are sampled from different sub-populations, which together form a distribution of population parameters. There are therefore two levels of sampling, and two sources of sampling error: within-study and between-study variation (Borenstein et al., 2009). The random effects pooled effect $\hat{\beta}_{RE}$ is calculated as the expected mean effect across this distribution of population effects, using a modified weighted average of the inverse of the variance incorporating the two sources of sampling error. Each study weight is equal to the inverse of the within-study error variance of the individual study s_i^2/n_i plus the estimated between-study variance τ^2, or $\frac{1}{s_i^2/n_i+\tau^2}$. Again, since the weight for a single study is equal to the inverse of the sum of the within and between study variances, the expected variance of the random effects average s_{RE}^2 is the inverse of the sum of the weights across the studies (Borenstein et al., 2009):

$$s_{RE}^2 = \frac{1}{\sum_i^k \frac{1}{s_i^2/n_i + \tau^2}} \qquad (4.9)$$

By making two further simplifying assumptions, that each study has the same population variance and sample size, it can be shown that the random effects variance is equal to:

$$s_{RE}^2 = \frac{\sigma^2}{kn} + \frac{\tau^2}{k} \qquad if\ n_i\ = n_k\ = n \qquad (4.10)$$

Hence the error variance is equal to the fixed effect (within-study) variance, which tends to zero as the study sample size increases, plus the estimated between-study variance, which tends to zero as the number of studies increases (Borenstein et al., 2009). As indicated by Hedges (1983), "[t]his model is appropriate when the studies used in the analysis are representative (if not a random sample) of a larger population and the researcher wants to generalize to that larger population" (p.389). The between-study variance can also be reduced by incorporating explanatory variables in meta-regression modelling, effectively attempting to capture those sub-population

characteristics that explain the between-study variation. The between-study variance can be estimated using the method of DerSimonian and Laird (1986):

$$\tau^2 = \max\left\{0, \frac{Q - df}{\sum_{i=1}^{k} w_i \quad - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}\right\},$$

$$where\ Q = \sum_{i=1}^{k} w_i \ \left(\hat{b}_i - \hat{\beta}\right)^2 \sim \chi^2_{df=k-1} \qquad (4.11)$$

where τ^2 is artificially constrained at zero if the value falls below zero (since a variance cannot be less than zero), and Q is the inverse-variance weighted sum of squares of the difference between treatment effects $\hat{b}_i$ and their estimated mean $\hat{\beta}$. Q is a statistic that follows the Chi-squared distribution with degrees of freedom $df = k - 1$, where Q represents the observed variation and df the expected variation based on sampling error alone. The denominator in the formula converts the difference $Q - df$ into units of the effect. Hence, the between-studies variance is measured as the estimated excess variation over that expected by sampling error, in the metric of the effect size (Borenstein et al., 2009).

A measure of the proportion of variance due to variation in the 'true' effects over sampling variation, I-squared, is calculated as (Higgins and Thompson, 2002; Borenstein et al., 2017):

$$I^2 = \frac{\tau^2/k}{s^2_{RE}} = \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2} \qquad (4.12)$$

under the assumption of equal study variance and sample size. I-squared is usually expressed as a percentage rather than a proportion.

A 95 percent confidence interval can also be calculated to show the uncertainty in the random effects average. However, there is additional uncertainty in whether the random effects average represents the population effect because of the estimated between-studies variance. The prediction interval calculates the confidence interval reflecting this greater uncertainty, calculated as (Riley et al., 2011):

$$\hat{b}_{RE} \pm t_{k-2}^{\alpha=0.05} \sqrt{s_{RE}^2 + \tau^2} \qquad (4.13)$$

where $t_{k-2}^{0.05}$ is the $100\left(1-\frac{\alpha}{2}\right)$ percentile of the t distribution with *k-2* degrees of freedom. It is interpreted as the interval in which the effect found in a new study will be incorporated, in 95 out of 100 cases (Masset, 2019).

It can be seen from equation 4.11 that the inclusion of estimators that deviate from the estimated mean effect due to bias, over and above the within- and between-study sampling error, will cause bias in the estimated between-study heterogeneity, pooled effect and I-squared (equation 3.12). It is therefore important to control for bias in estimation, which is usually done through critical appraisal. For example, evidence from meta-analyses of education programmes in low- and middle-income countries suggests NRS with credible means of control for confounding can produce the same pooled effects as RCTs (Table 4.1). NRS included in the education meta-analyses used difference-in-differences, instrumental variables, propensity score matching and regression discontinuity design (Baird et al., 2013; Petrosino et al., 2012).

Importantly, the evidence presented in Table 4.1 suggests that, where there is greater scope for self-selection into intervention group and/or selective reporting of outcomes, as in the case of microcredit (Vaessen et al., 2014), NRS are more likely to estimate larger treatment effects than RCTs, which may suggest bias. There is arguably greater risk of self-selection into microcredit groups, and subsequent receipt of loans, than there is of self-selection into cash transfers or education interventions, where decisions about who should participate in intervention are taken by programmers. In addition, household spending decisions were largely reported, whereas many enrolment and attendance outcomes were observed, which may introduce further bias in microcredit evaluations.[49] Hence, the pooled effects from NRS on microcredit deviate more from the RCT estimate, than either cash transfers or education.[50]

[49] As noted in Vaessen (2014, p.39): "[s]tudies generally collected self-reported outcomes from survey questionnaires over a range of expenditure items which were grouped into a composite index". In contrast, although some studies used self-reporting by the household in Baird et al. (2014), others used unannounced school visits by researchers.

[50] It may also be of interest to know whether self-selection (which can be addressed through improved study design) or selective reporting of outcomes (which can be

Table 4.1 Pooled effects of RCTs and NRS of interventions in L&MICs

Outcome	*Design (bias)*	*OR*	*95% CI*		*P>|z|*	*Tau²*	*I²*	*MSE+*	*obs*
Enrolment*	RCT	1.40	1.21	1.61	0.000	0.06	90%	0.065	15
	NRS	1.38	1.25	1.52	0.000	0.04	87%	0.043	27
Attendance**	RCT	1.33	1.20	1.46	0.000	0.02	91%	0.023	43
	NRS	1.34	1.20	1.52	0.000	0.02	97%	0.024	16
Woman makes household spending decisions***	RCT	0.99	0.93	1.05	0.437	0.00	0%	0.001	4
	NRS ('some concerns')	1.04	0.90	1.20	0.064	0.00	64%	0.008	3
	NRS ('high risk of bias')	1.16	0.98	1.36	0.000	0.02	86%	0.052	11

Notes: + MSE uses the natural logarithm of *OR* and its standard error; it is calculated for RCTs assuming bias=0. *OR* estimated by inverse-variance weighted random effects meta-analysis. Interventions are * cash transfer versus control (Baird et al., 2013), ** education intervention versus standard intervention (Petrosino et al., 2012) and *** access to microcredit versus control (Vaessen et al., 2014).

Source: author based on reported data.

In addition, systematic reviews have different inclusion criteria, and reviews with broader study design inclusion criteria are more likely to produce biased pooled effects. In this case, the review on microcredit included many *a priori* less credible studies, in particular those applying adjusted regression analysis to post-test cross-sectional data (Vaessen et al., 2014). In contrast, the review on education excluded any study without pre-test measurement (Petrosino et al., 2012). And while the review of cash transfers incorporated studies using cross-sectional data, the NRS evidence base largely consisted of studies with more credible methods of analysis such as DD, RDD and statistical matching (Baird et al., 2014). When the NRS in Vaessen et al. (2014) were separated into high and medium risk of bias,[51] where medium risk studies all used identification methods thought to be more internally valid (RDD, IV or statistical matching), the pooled estimate of the 'medium risk of bias' studies was closer to the RCT estimate (Table 4.1).[52] But it was

addressed through improved outcome data collection) are the critical factors in determining bias.

[51] Determining overall risk of bias is complicated because the degree of bias is a latent construct (i.e., one that is not directly observable or measurable). However, it is useful as shown in this and the following chapter (see also Guyatt et al., 2011).

[52] No NRS (or, for that matter, RCTs) in the review were identified by the authors as having low risk of bias. The risk of bias assessments used in Baird et al. (2014)

still not as accurate as in the case of cash transfers and education, suggesting residual confounding due to self-selection of participants to microcredit groups and receipt of loans.[53]

However, there are other threats to validity in making generalisations across studies, due to systematic factors that affect the distribution of observed effects. One is sampling bias; another is publication bias. Publication bias is usually thought to cause lower censoring of the distribution of effects. There are standard approaches to attempt to deal with the problem, including searching for unpublished studies, the assessment of reporting biases in critical appraisal (see below Section 4.4.5), and statistical testing based on small-sample bias (Egger et al., 1998; Peters et al., 2008).

Addressing sampling bias is more difficult. In impact evaluation, there is usually no clearly defined (sub-) population to which the results are expected to generalise (Tipton, 2013). One argument is that as the number of studies increases, so does the likelihood that the studies are representative of the population (Borenstein et al., 2009). Methods such as meta-regression modelling can also attempt to account for non-randomness in the distribution of effects. Some authors apply meta-regression modelling alongside Bayesian meta-analysis in the attempt to estimate more accurate pooled effects. For example, Vivalt (2020) aims to answer the question 'how much can we generalize from impact evaluations?'. In contrast, Tipton (2013) proposes an approach using propensity score matching to generalise the findings from one study to another context. At the very least, it would seem to provide further grounds for greater care in interpreting random effects meta-analysis and therefore the use of prediction intervals as standard.

4.3 Categorising impact evaluations

Impact evaluations are usually, implicitly, characterised by the extent to which they can address confounding by design or in analysis. Confounding

and Vaessen et al. (2014) use the approach by the author (Hombrados and Waddington, 2012), which is further developed in this chapter.

[53] Using the distance metric defined in Chapter 4 equation 4.7 below, the absolute standardised mean difference is 0.099 for cash transfers and 0.075 for education. Whereas in the case of microcredit, it is 0.796 for medium risk of bias NRS, and 2.560 for high risk of bias studies.

can be observed or unobserved (unmeasured or unmeasurable), time-invariant (fixed over the course of the study at baseline) or time-varying. For example, confounders in the relationship between access to latrines and reported diarrhoea include: readily observable factors like sex and age; more complex factors like socioeconomic status, which can be measured imprecisely using wealth indices in DHS (Filmer and Pritchett, 2001), or approximated through expensive household income and expenditure surveys; factors that are often unmeasured such as hand hygiene practices or the degree of functioning and use of water supply (Cairncross and Kolsky, 1997); and factors which are arguably unobservable such as self-efficacy, attitudes to risk, behavioural responses to incentives by research participants (e.g., bias in self-reported outcomes) (Schmidt and Cairncross, 2009). Some of these confounders are usually fixed or time-invariant throughout a study or baseline values can be readily recalled (e.g., sex, age); others are more likely time varying (e.g., functioning of infrastructure, behaviour change in response to interventions, self-efficacy). Confounders can also be differentiated from mediators, which are intermediate factors along the causal pathway such as latrine functioning and use, and exposure to environmental contamination via open defaecation (Table 4.2).

Table 4.2 Variables affecting the observed effect of latrine access

Type	*Observable confounders*	*Unmeasured confounders*	*Unobservable confounders*	*Mediator variables*
Example	Sex Age Location Assets Functioning of water supply	Hand hygiene behaviour Use of water supply Socioeconomic status	Self-efficacy Attitude to risk Behavioural response to incentives (e.g., agreeableness)	Latrine functioning Latrine use Open defaecation

Source: author.

Some types of confounding bias can be controlled in analysis. For example, observables can be controlled in adjusted analysis, assuming they can be measured precisely; time-invariant confounding (including unobservables) can be controlled through statistical modelling where pre-test post-test outcomes data are available (e.g., double differences). However, unobservable confounders, which are more likely to be measured at the individual level, can most effectively be controlled in study designs which are able to control for unobservable and observable confounders where factors

determining allocation to intervention are precisely known (e.g., RCTs and RDDs). In these studies, the "control group provides an unbiased estimate of the average *potential outcome* [(Rubin, 1974)] that experimental units would have attained had the treatment not been applied to them" (Cook and Steiner, 2010, p. 57).

It would also seem intuitively reasonable that confounding due to factors determining programme placement at group level (called 'programme placement bias') may be easier to observe – and therefore control – than confounding due to self-selected uptake or adherence (participant 'self-selection bias').[54] Confounding due to self-selection is thought more problematic in studies of latrine provision than water supply provision, simply because individuals within a community tend to self-select to install their own latrine, whereas water supply tends to be provided by the public agency to the community as a whole. For example, Hoque et al. (1995) in Bangladesh and Strina et al. (2003) in Brazil found households with latrines were significantly more likely to undertake other improved behaviours like hygiene. Furthermore, when programmes are geographically targeted, there is likely to be greater unobservable confounding across locations than within them, complicating evaluation design (Handa and Maluccio, 2010). These may underlie Cook et al.'s (2008) finding that statistical matching is more accurate when it is done of intact clusters rather than of individual cases, since it may be difficult to identify suitable matches for individual cases across clusters (e.g., to account for spillover effects or contamination). If a programme is rationed by supply, such as installation of a village handpump or connection of latrines to the public sewerage network, information is needed on the criteria determining rationing (e.g., a threshold, geographical characteristics, socio-demographic or economic factors). In contrast, where a programme is demand-driven, individual characteristics determining participation must be understood, which are likely to be difficult to observe or model.

Information about the programme targeting approach may therefore be particularly useful in formulating strategies to approximate the (usually unobserved) selection process in non-randomised studies (e.g., Campbell, 1984; Cook et al., 2008). Targeting mechanisms can be divided into three

[54] Note, this is different from 'sample selection bias', which is referred to as 'selection bias' below in Section 4.4.2.

broad types (Coady et al., 2003). 'Individual/household assessment' involves either a means test or the selection of participants according to explicit criteria by a third party such as community leaders or programme implementers. 'Categorical' targeting identifies target groups using easily identifiable criteria at either the individual or household level (e.g., gender, age, ownership of land, membership of farmer group), or the community level (e.g., specific locations, areas with pest or pesticide problems). 'Self-selection' occurs where a programme is universally available. Furthermore, the specific targeting criteria for groups or individuals can be categorised into those that may favour successful implementation and effectiveness (e.g., localities with strong existing community groups, individuals selected to participate due to social standing or resources like land), those favouring equity or inclusion (e.g., of women, poor, elderly or disabled), factors relating to exposure to infectious diseases (likely combining effectiveness with equity), and practical criteria relating to convenience, accessibility and availability (Box 4.1).

Study designs for causal inference differ according to the extent to which, when well implemented, they can address observable and unobservable confounders. Some account for unobservable confounding by design, either through knowledge about the method of allocation or in the methods of analysis used. These designs, termed 'selection on unobservables', include RCTs, natural experiments, regression discontinuity designs (RDDs) and studies using instrumental variables or double differences estimation (Imbens and Wooldridge, 2009). Other studies can address selection on observables only, including non-randomised studies that control directly for confounding in adjusted analysis (e.g., single difference studies using statistical matching, analysis of covariance, multivariate regression). These studies assume 'unconfoundedness', a property that is unverifiable, although falsification tests exist (e.g., Rosenbaum and Rubin, 1982). Studies using double differences (e.g., difference-in-differences, triple differences) and fixed or random effects regression analysis using panel data with measurement of outcomes at pre-test and post-test, are intermediate cases, where unobservable confounders that are fixed over time can be controlled at the unit of analysis.[55] An example may be household hygiene behaviour in

[55] The existence of time-varying unobservables may be assessed by comparing parallel trends in the outcome (double differences) or estimating a leads and lags model (fixed or random effects)

the case of a water or sanitation infrastructure programme (assuming there is no contemporaneous hygiene behaviour change campaign).

Box 4.1 Programme targeting mechanisms and criteria

Mechanisms

- Categorical/group-based: all individuals in a specified category are eligible such as selected communities, geographical locations, demographic characteristics (e.g., age group or sex) or socioeconomic factors (e.g., land ownership).
- Individual/household assessment: those eligible according to a proxy-means test (e.g., asset index), selected by practitioners, or by the community.
- Self-selection: eligibility is universal, but benefits may be provided in such a way as to encourage uptake by desired groups and discourage uptake by others (e.g., service delivery points like water pumps are located in areas where poor people are concentrated).

Effectiveness criteria – target those considered most able to make best use of the WASH technology

- Resources: only those with access to some land or water supply.
- Social standing: those with social standing/influence.

Equity criteria – target those considered to be most in need

- Women: designed to benefit women and children.
- Pro-poor: landless, marginal, poor or those with few resources.
- Inclusivity: intended to include those who are vulnerable (e.g., young children, elderly, HIV affected) or disadvantaged (e.g., by education, resource or socio-economic level).

Combined equity and effectiveness criteria

- Disease: households or communities with known exposure to infectious disease.
- Pre-existing groups: e.g., community groups, women's health clubs.

Practical criteria

- Accessibility: localities chosen for accessibility, proximity to roads or water source, or chosen because of existing development operations.
- Convenience: households located close to one-another.
- Availability: individuals available and with time to participate.
- Interest: individuals motivated and interested in participating.

Source: adapted from Phillips et al. (2014); Coady et al. (2003).

Study designs can also be differentiated according to whether they are designed prospectively at pre-intervention stage, or retrospectively designed post-intervention. These categories are usually synonymous with whether

the study is experimental[56] – that is, the intervention and data collection are centrally controlled, usually by the investigator – or observational – where the intervention (and often the data source) are independent of research investigation (Shadish et al., 2002). Craig et al. (2011, p.7) further differentiated natural experiments as studies where there is "unplanned variation in exposure" to intervention which is used to attempt to make causal inference.

Dunning (2012) is more specific, characterising natural experiments as those applying statistical techniques, often to observational data sets, using knowledge about natural processes of programme assignment (e.g., policy, geography) to generate as-good-as randomised ('as-if randomised') assignment. According to Dunning (2012), therefore, these are retrospective observational studies with selection on unobservables.[57] Purely observational studies are retrospective studies of observational data with selection on observables only, where treatment decisions are made by self-selection of participants, practitioners or planners. Quasi-experiments therefore comprise the remaining non-randomised studies that are prospective in design, where measurement is centrally controlled by investigators for the explicit purpose of evaluating the intervention of interest, and where the investigators may have some control over scheduling treatment and selecting comparison groups, even if treatment itself remains self-selected (as it does in all voluntary programmes).

Table 4.3 shows this classification of research designs according to four questions: 'is the research undertaken prospectively?'; 'is treatment centrally controlled (e.g., nature and timing of treatment and dosage)?'; 'are units of analysis randomly allocated?'; and 'is measurement centrally controlled (e.g., who is measured, on what, when, how often)?'. As we will see later, this classification is useful because it helps inform potential threats to internal validity in critical appraisal analysis, especially in differentiating threats due to confounding and selection bias (usually more problematic in observational designs) from observer and responder bias (more problematic in trials) (Schmidt, 2014). The table also shows the classification of impact

[56] Shadish et al. (2002, p.12) define an experiment as "[a] study in which an intervention is deliberately introduced to observe its effects."
[57] Dunning (2010) also further differentiates 'randomised natural experiments', where there is randomisation by policy-makers to a condition by a lottery process (e.g., Vietnam war draft), from other natural experiments.

evaluations in the WASH evidence census. Of the prospective NRS, called quasi-experiments here, 90 percent were done using data collected by the authors, the remaining 10 percent using existing data. For natural experiments, the opposite was found (only 10 percent collected own data).

Table 4.3 Classifying research designs for causal inference

	Prospective?	*Treatment centrally controlled?*	*Random assignment?*	*Measurement centrally controlled?*	*Num. WASH studies*
Experiment (RCT)	Y	Y	Y	Y	225
Quasi-experiment (prospective NRS)	P	P	N	Y	115
Natural experiment ('as-if' randomised retrospective study)	N	N	Y*	P	11
Observational study (retrospective NRS)	N	N	N	N	13

Notes: 'Y' yes; 'P' potentially; 'N' no; * 'as-if' random via natural variation.

Source: adapted from figure provided by Scott Bayley (pers. comm.).

There are numerous examples of the use each design in WASH evaluations. Since it defined the present field of study, let us look in detail at John Snow's study of cholera (Snow, 1855), a compendium of three investigations into cholera outbreaks in London 1849, 1853 and 1854. Snow presented many examples to support his belief that cholera transmission was largely water-borne. Many of these fulfil Bradford-Hill's (1965) criteria for determining a causal relationship: strength of association (effect size), consistency in evidence, specificity, temporality, biological gradient (dose response), plausibility, coherence, experimentation and analogy (Table 4.4). Bradford-Hill's criteria are often used in WASH impact evaluations to support inferences made about attribution, especially in NRS and in the presence of implementation errors in RCTs. For example, falsification tests for the specificity of causal pathways may be made with reference to 'negative controls' such as a non-equivalent independent variable function (also called a 'placebo intervention') – that is, a concurrent intervention received by study participants that is unrelated to outcomes of interest – or a non-equivalent dependent variable function ('placebo outcome') – an outcome measured among treatment groups that is unrelated to interventions of interest (Lipsitch et al., 2010). Indeed, several of these criteria are used in appraisals of the body of evidence in systematic reviews and meta-analyses

using GRADE (Guyatt et al., 2011): magnitude of effect size, dose response and consistency.[58]

Snow presented administrative data on death rates from cholera outbreaks in 1849 and 1853 in districts of London that received water supply from different companies. Noting that companies obtained water from the Thames at different points, specifically that the Lambeth Water Company, previously taking water from the Thames downstream from the sewage outlet, had moved its intake upstream in 1852, compared to the Southwark and Vauxhall Company which had kept the intake source downstream, Snow shows, in a controlled before-versus-after (CBA) observational design using administrative data, that districts receiving water supply solely from Southwark and Vauxhall had higher rates of cholera mortality than those receiving water from both Southwark and Vauxhall and Lambeth Water Companies, whereas those supplied by Lambeth alone had no cases. However, this would not count as incontrovertible evidence due to the other sources of potential confounding of possible transmission routes in district level data, such as poverty and population density.[59]

However, in what has been called a natural experiment (Dunning, 2012) and a quasi-experiment (Bärnighausen et al., 2017a),[60] Snow observed that the nature of the competition in the market for water supply meant that water pipes from different water utility providers went "down all the streets, and into nearly all the courts and alleys" (1855, p.74).

Snow provides a useful description of the benefits of ('as-if') randomisation:

[58] Two criteria used in GRADE are tangentially related: bias relating to experimentation, and indirectness relating to specificity. Two further criteria included in GRADE are additional: precision (statistical significance) and publication bias (systematic bias in reporting).
[59] At the time, another theory about cholera transmission was that it spread via "effluvia given off from the patient into the surrounding air, and inhaled by others into the lungs" (Snow, 1855, p.9).
[60] According to the schema presented here, the study classifies as a hybrid natural quasi-experiment. It has elements of quasi-experiment, because it was designed prospectively, and data were collected by Snow for the purposes of the study. It appears that Snow was not able to conduct the cohort analysis during the cholera epidemic in 1853, and so waited until the epidemic of the following year to collect the data (Snow, 1855). However, it is a natural experiment because the process determining 'as-if' randomised treatment assignment was outside of Snow's control. Taking the definition of natural experiment from Craig et al. (2011), the study is classifiable as a natural experiment.

> "As there was no difference whatever, either in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.
>
> "The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewerage of London, and, amongst it, whatever might have come from cholera patients, the other group having water quite free from such impurity."
>
> Snow (1855, p.75).

Snow correlated administrative data on cholera deaths with water supply source, obtained by interviewing households at, and collecting water samples from, addresses where known cholera deaths occurred. The result of investigation, presented in Table 4.4, showed a big and precisely estimated odds ratio (OR) of 13.32 (95% confidence interval, 95%CI=7.84, 22.93)[61] higher cholera deaths in households supplied by sewer-contaminated water, versus those not. Or alternatively, death rates among those living in households with uncontaminated water supplies were 92 percent lower than those in households with contaminated water during the cholera epidemic (OR=0.08, 95%CI=0.04, 0.13).

[61] Author's calculation assuming independence of observations.

Table 4.4 Criteria for determining cause from association

Criterion	*Definition from Bradford-Hill (1965)*	*Examples from 'Snow on Cholera' (Snow, 1865)*
Strength	Strength of association between exposure and outcome.	According to Snow, in the 1849 epidemic there were 856 cholera deaths in 77,796 living (11 per 1,000) in Southwark, where water was supplied without filter or settling reservoir, compared to 325 deaths in 124,585 living (2.6 per 1,000) in Westminster, supplied by a company using settling reservoirs and filters. In the 1854 epidemic there were 286 deaths in 40,046 houses (7.2 per 1,000) under Southwark and Vauxhall Company versus 14 deaths in 26,093 houses (0.5 per 1,000) in Lambeth Company.
Consistency	Association is observed in different times, places, circumstances, by different persons, and different methodologies (e.g., prospectively and retrospectively).	Snow referred to documented cholera outbreaks in 1832 in Newburn, England, 1814 in Cunnatore, India, the Baljik Bay, now Bulgaria, due to contaminated water supply. Examination of the 1849 cholera outbreak in Broad Street, north London, and 1853 and 1854 outbreaks in south London, related populations drinking contaminated water supply using case report, controlled before-and-after and (natural) experiment.
Specificity	Association is specific to particular causal pathways and there is no association between the exposure and other (irrelevant) outcomes, or the outcome and those not exposed to the cause.	Snow observed of the 1849 epidemic, a Workhouse on nearby Poland Street was surrounded by houses in which deaths from cholera occurred but only 5 deaths in 535 inmates occurred, all of whom were admitted after contracting cholera. "The workhouse has a pump-well on the premises… the inmates never sent to Broad Street for water" (p.42). There was also a brewery in Broad Street, near the pump, where no men were confirmed as having cholera, at least severely. "The men were allowed a certain quantity of malt liquor, and [the proprietor] believes they do not drink water at all; and he is quite certain that the workmen never obtained water from the pump in the street. There is a deep well in the brewery" (p.42).
Temporality	Cause must precede effect on the outcome.	According to Snow: "In cholera, [the] period of incubation or reproduction is much shorter than in most other epidemic or communicable diseases. From the cases previously detailed, it is shown to be in general only from twenty-four to forty-eight hours" (p.16). Snow observed of the 1849 epidemic: "The first case of decided Asiatic cholera in London, in the autumn of 1848, was that of a seaman… who had newly arrived… from Hamburgh, where the disease was prevailing… He was seized with cholera on the 22nd of September and died in a few hours. Now the next case of cholera, in London, occurred in the very room in which the above patient died… He was attacked with cholera on the 30th September." (p.3)

Criterion	*Definition from Bradford-Hill (1965)*	*Examples from 'Snow on Cholera' (Snow, 1865)*
Biological gradient	The association between exposure and outcome reveals a dose-response relationship.	Of the 1849 epidemic, Snow observed a positive association between proximity to the Broad Street well and deaths due to cholera: "deaths either very much diminished, or ceased altogether, at every point where it becomes decidedly nearer to send to another pump than the one in Broad Street" (p.47). Cholera also propagated more in the "crowded habitations of the poor, in Westminster [where there were 6.8/1,000 deaths from cholera], than in the commodious houses of the Belgrave district [2.8/1,000 deaths]" (p.66 [data from Table III pp.62-63]). Of the 1853 epidemic, there were 11.4 cholera deaths per 1,000 population in the districts that were solely supplied by contaminated water, 6 per 1,000 in districts supplied by some contaminated and some uncontaminated sources, and zero deaths in districts supplied solely by uncontaminated sources (p.73 Table VI).
Plausibility	The causation is theoretically plausible (although what is plausible depends on the scientific knowledge of the day).	Snow believed cholera to be water-borne and communicated from person to person through contact with bodily secretions, rather than through airborne transmission. On the mode of communication, he noted: "Nothing has been found to favour the extension of cholera more than want of personal cleanliness, whether arising from habit or scarcity of water… The bed linen becomes wetted by the cholera evacuations, and these are devoid of the usual colour and odour, the hands of the persons waiting on the patient become soiled without their knowing it; and unless these persons are scrupulously cleanly in their habits, and wash their hands before taking food, they must accidentally swallow some of the excretion, and leave some on the food they handle or prepare, which has to be eaten by the rest of the family, who, amongst the working classes, often have to take their meals in the sick room: hence the thousands of instances in which, amongst this class of the population, a case of cholera in one member of the family is followed by other cases; whilst medical men and others, who merely visit the patients, generally escape." (p.16-17) In addition, Snow noted that "[f]or the morbid matter of cholera having the property of reproducing its own kind, [it] must necessarily have some sort of structure, most likely that of a cell. It is no objection to this view that the structure of the cholera poison cannot be recognised by the microscope, for the matter of small-pox and of chancre can only be recognised by their effects, and not by their physical properties." (p.15).

Criterion	*Definition from Bradford-Hill (1965)*	*Examples from 'Snow on Cholera' (Snow, 1865)*
Coherence	The causation does not seriously conflict with other known facts about the outcome and how it occurs.	Snow observed of the 1849 epidemic: "The only other water company deriving a supply from the Thames, in a situation where it is much contaminated with the contents of the sewers, was the Chelsea Company. But this company… took great pains to filter the water before its distribution" (p.64). There were 2.8 deaths per 1,000 from cholera in areas covered by Chelsea water supply, as compared to up to 21.5/1,000 in areas covered by Southwark and Vauxhall Company. In addition, Snow noted that "[a]s cholera commences with an affection of the alimentary canal, and as we have seen that the blood is not under the influence of any poison in the early stages of this disease, it follows that the morbid material producing cholera must be introduced into the alimentary canal – must, in fact, be swallowed accidentally, for persons would not take it intentionally; and the increase of the morbid material, or cholera poison, must take place in the interior of the stomach and bowels. It would seem that the cholera poison, when reproduced in sufficient quantity, acts as an irritant on the surface of the stomach and intestines, or, what is still more probable, it withdraws fluid from the blood circulating in the capillaries, by a power analogous to that by which the epithelial cells of the various organs abstract the different secretions in the healthy body." (p.15)
Experiment	By manipulating the cause, it should be possible to change the frequency of associated events; "the strongest support for the causation hypothesis may be revealed" in this way (pp.298-9).	During the 1854 epidemic, Snow conducted a study of streets covered by water pipes from both Southwark and Vauxhall Company and Lambeth Water Company, asking households to identify the water company providing their source. When they could not answer, he was able to determine the source that each house received using chemical test, due to the "great difference in the quantity of chloride and sodium contained in the two kinds of water" (p.78) from the two water companies. Using administrative data, he was able to correlate cholera mortality in households provided by each source.
Analogy	Evidence of a causal relationship between similar exposures and the outcome is acceptable in some circumstances.	Snow noted that "[t]here is a good deal of evidence to show that… typhoid fever, and yellow fever, diseases in which [like cholera] the blood is affected, are propagated in the same way as cholera" (p.16).

Sources: Snow (1855) and Bradford-Hill (1965).

As noted in Chapter 1, in 1927, Stockton-on-Tees, England, "favourable circumstances for human field research" (M'Gonigle and Kirby, 1937, p.109) enabled quasi-experimental investigation of the effects of improved housing on nutrition. The research was designed prospectively, with "arrangements made to keep careful records of" (p.108) two population groups. Owing to the phased roll-out of slum clearance and rehousing, a treatment group comprising 152 families and 710 individuals was transferred to a self-contained housing estate, while a comparison group comprising 289 families and 1,298 individuals remained in slum housing. The authors found that, standardising by age and sex distributions of the two populations, death rates in the treated group were observed to fall from 34 to 23 per 1,000 living, while they rose in control group from 23 to 26 per 1,000.[62]

This study was an early example of cross-section evaluation design. Cairncross et al. (1980) differentiated four main types of water project evaluation design according to the groups enrolled and data collection periods. These include: (a) the 'ideal type' with both pre-test (baseline) and post-test (follow-up) data among an intervention group and separate control group; (b) 'cross-section surveys' with post-test data collection only; (c) 'time series study' with pre-test and post-test data collection in intervention group only; and (d) 'case study' with post-test data collection in intervention group only. They stated, with uncharacteristic pessimism, that "unless the design is of the form of (a)… there are severe impediments to attributing any observed changes to the improved water supply" (p.11).

The earliest controlled impact evaluations of WASH improvements in L&MICs were of water supply improvements (Feachem et al., 1978), sometimes alongside domestic hygiene education (Shiffman et al., 1978). These studies tended to be done in a few villages, with data collected from multiple households within each village. They often lacked the sample sizes

[62] Given the higher death rate in treated group at baseline, it is likely that this population was moved by the authorities first due to greater need. These represent pre-existing differences that would invalidate simple non-randomised comparisons. Interestingly, the authors note "[f]or convenience a line of division was decided upon which ran along a street called 'Smithfield'" (p.108), which demarked the treatment and comparison groups. Had it been possible to follow up households who moved from Smithfield Street, and the sample size large enough for statistical precision (unlikely for mortality due to rarity of observation, but possible in theory for other outcomes data collected like food purchases), it may have been possible to accurately measure the effect of improved housing by comparing these households with those who remained living in Smithfield Street, using geographical discontinuity design (GDD) (see Section 4.4).

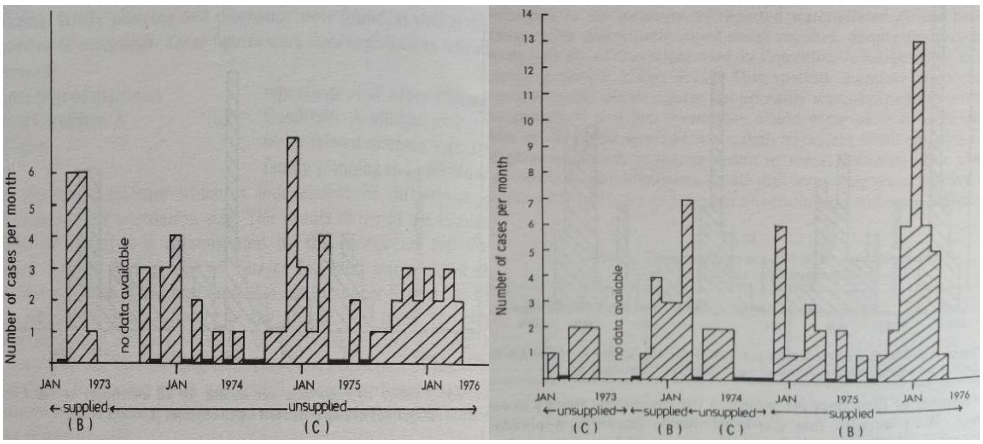
to estimate effects with statistical precision, due to intra-cluster correlations of observations within villages. In Lesotho, Feachem et al. (1978) observed changes in time use and water use quasi-experimentally in cross-section design, following water supply improvements (e.g., from unprotected spring or waterhole to protected spring with storage and reticulation or borehole with hand pump) in 58 villages, estimating time savings of 30 minutes on average per day per adult women. The authors linked administrative records from observational health facility diarrhoeal disease records (including typhoid), with natural variation in village water supply characteristics. They divided villages into four groups over the three-year period of study: those never having improved supplies; those always having improved supplies; and intermediate cases, those with an improved supply that worked most of the time, and those with an improved supply that was broken down most of the time. Observing the peaks in diarrhoeal disease were in the wet season in all cases, regardless of the quality of water supply, they argued that disease transmission was largely water-washed rather than water-borne: if it were water-borne, disease would have been more prevalent in the wet season where people used unimproved sources that were contaminated by faeces; whereas the incidence did peak in wet season but was bigger in villages with improved sources (protected springs and boreholes) that could not be contaminated in this way.

The intermediate cases may form a natural experiment, where periodic breakdowns unrelated to village assignment created exogenous variation in access to improved water supply in some villages, which remained subject to the same peaks in diarrhoea during the wet season (Figure 4.1).[63] It would otherwise be difficult to argue that villages with and without improved supplies were equivalent.[64] An additional advantage of using contemporaneous health seeking outcomes from health facility records, is that selection biases due to losses to follow-up (of eligible treatment units or follow-up periods) would have been minimised over the course of the study.

[63] In another natural experiment, Zafar et al. (2015) examined surgery success according to hour of day.

[64] Feachem et al. (1978) argued that the villages with improved water supplies were "in effect a random selection and are statistically comparable to those which have not" (p.181). This was based on examination of characteristics (e.g., time spent collecting water prior to installation of improved supply) and knowledge about the political decision-making process, since nearly all villages had applied for water supply improvements, and the ruling party did "not use distance to the source or other such criteria when selecting the villages which are to receive supplies from the list of villages which have made applications" (p.182).

Figure 4.1 Diarrhoea reports per month from two villages in Lesotho with improved water supplies subject to periodic breakdown



Source: Feachem et al. (1978).

In Mozambique 1982, Cairncross and Cliff (1987) conducted a pipeline quasi-experimental cross-section evaluation of time allocation for women living in two villages in northern Mozambique. Water supply in *Namaua* village was a standpipe on average 300m away from each household. The standpipe had been provided by government prior to the evaluation and was in good working order. For residents of *Itanda*, which was due to receive its own standpipe shortly after, water supply was available in a neighbouring village 4km away. Data were collected by observing adult females on two consecutive dates in each village. They found significant time savings, on average nearly two hours per day per woman, which were largely spent doing household work (e.g., food preparation and childcare), personal care (e.g., hygiene) and income generation, although the latter was not statistically significant (Table 4.5). This suggested that improved water supply may contribute to child nutritional status through the following proximate determinants (as later classified by UNICEF, 1990): household food security, exposure to infections such as diarrhoea (via the water-washed route), and quantity and quality of childcare (determining how effectively income is converted into nutrition and the share allocated to children).

The authors were effectively approximating a reflexive control (pre-test post-test) design, without having to rely on inaccurate recall, by using a pipeline design (that is, estimating the effect with reference to another village that is as similar as possible to the treatment village, including by being eligible for future treatment). They presented some information suggesting that the

comparison is valid, such as the average time use in each village being statistically identical, and the eligibility for treatment of the unsupplied village suggesting confounding due to programme placement bias may not be problematic.

Table 4.5 Average time budgets for the observed waking day of adult women (in minutes); Mueda, Mozambique

Activity	*Unimproved water supply (n=110 women-days)*	*Improved water supply (n=118 women-days)*	*Difference**
Water collection including queueing time	131	25	-106 (-135, -77)
Other household work including food preparation, childcare	126	161	35 (6, 64)
Grinding cereals	84	98	14 (-15, 43)
Agricultural work	154	160	6 (-23, 35)
Rest including time for eating, personal hygiene, education	384	433	48 (20, 78)
Total	880	877	-3 (-32, 26)

Note: * 95 percent confidence intervals in parentheses calculated by author from reported standard deviation (150 minutes), assuming independence of observations.
Source: Cairncross and Cliff (1987).

However, the authors used what might be called 'naïve matching' rather than statistical matching, and other reported factors suggest the comparison is imperfect, such as the differences in village size (2,800 people in Namua and 1,200 in Itanda). Efforts were made to ensure quality of data, such as by collecting outcomes data through observation rather than self-report. An alternative design, as noted by the authors, would have been to collect data on women bathing and washing using surface water in Namaua, presumably subject to even greater confounding due to self-selection. However, stronger

inference may have been possible if pre-existing observable characteristics of villages were compared directly, and differences between individuals controlled in adjusted analysis, as well as more villages included in each study arm to increase the effective sample size (see below Section 4.4.6).[65]

As noted in Chapter 3, Section 3.4.1, there has been a big increase in the number of large-sample, prospective impact evaluations of WASH interventions in L&MICs. Early randomised controlled trials (RCTs) were done of household level interventions, for example of household water treatment in the Gambia (Austin, 1993) and Guatemala (URL, 1995), water storage containers in a Malawi refugee camp (Roberts et al., 2001), household water treatment in Pakistan (e.g., Luby et al., 2004), Bolivia (Clasen et al., 2004) and Ethiopia (Boisson et al., 2009), and hand hygiene in Pakistan (Luby et al., 2004). Cluster-RCTs are increasingly commonplace, to examine interventions delivered at group level, such as source water protection in Kenya (Kremer et al., 2011), latrine provision in India (Clasen et al., 2014; Patil et al., 2014), and sanitation promotion in Mali (Pickering et al., 2015). Cluster-RCTs have also been done of treatments provided at household level, for example of household drinking water treatment in Bolivia (Mäusezahl et al., 2009).

There are several criteria determining whether variation in implementation of a programme can enable a control group to be identified: oversubscription or rationing of resources, phase-in of programmes over time (pipeline), within-group randomisation (use of 'active control'), and encouragement design (Duflo et al., 2006). White (2013) also refers to raised threshold design, where the programme admission criteria are extended so that controls can be identified who would otherwise be eligible, and factorial design, where different treatment combinations are compared to one another individually and together as co-interventions, against a control. Methodological developments like randomised encouragement enable more rigorous evaluations of interventions that would be difficult or impossible to conduct under pure controlled conditions. For example, where programme eligibility is universal, so a pure controlled study design is not possible, but programme take up is less than universal, programme marketing

[65] The correlation between observations within each village (intra-cluster correlation) would be expected to be smaller for socioeconomic outcomes than infectious diseases, although observations are still correlated since individuals in the same community use the same water sources (Cairncross et al., 1980).

information can be randomly assigned to treatment groups. For example, randomised encouragement was used in the evaluation of a programme providing credit to households for piped water connections in urban Morocco (Devoto et al., 2012).

However, random allocation of the treatment (or encouragement) is not always possible. Rubin (1974) and Attanasio (2011) give a series of circumstances that make experimental approaches to evaluation impossible. Firstly, RCTs may be prohibitively expensive.[66] Secondly, RCTs may not be ethical for some interventions or outcomes, such as the impact of smoking on lung cancer, or of WASH programmes on diarrhoea mortality. Thirdly, RCTs may be inappropriate in evaluation of long-term outcomes or of universal policy interventions when the change affects the whole population. Finally, RCTs are not possible for *ex post* evaluations, in which the treatment has been already assigned, or where policy makers want to use non-random targeting rules and there are insufficient observations to randomise among target groups.[67]

Moreover, some factors potentially diminish the internal validity of the approach for socioeconomic interventions. Deaton (2010) argues that specific technical problems arise in implementation of RCTs due to the impossibility of double blinding (of participants and investigators) to intervention, leading to imperfect compliance. This occurs especially when the treatment requires a behaviour that the participant is unwilling to undertake (e.g., use of a latrine or household water treatment device), where interventions are 'sustained' (that is, they require sustained adherence). In such cases, one might expect relatively high and non-random non-compliance of participants, also called 'no-shows', so that those that end up adhering to the intervention are not a random sample of the population

[66] The cost of a prospective study is primarily due to the costs of data collection, which is itself a function of the sample size, number of data collection rounds, type of data collected (e.g., whether reported, or observed and verified by laboratory testing), and competition in the market for survey organisations. For example, in sub-Saharan Africa where survey organisations are relatively few, typical costs of impact evaluation were up to US$ 1 million. In South Asia, where there are more survey organisations, costs were typically US$ 0.5 million. However, due to their greater statistical efficiency, randomised designs are likely to be less costly than prospective non-randomised studies (White, 2014).

[67] King (2009, p.487) argued that "[w]hen decisions are recognized as arbitrary, randomizing those decisions becomes acceptable. Because some decisions are always made below the level of political radar… randomization is always acceptable at one level below that at which politicians care."

initially assigned to the treatment group. In other words, another form of confounding due to self-selection, selective compliance, could cause the individuals in the treatment group to be incomparable with individuals in the control group, with the risk that the differences in outcomes might be explained by other factors (e.g., unobservable characteristics such as self-efficacy or attitudes towards risk) rather than participation. The problem may be corrected using instrumental variables estimation (see below Section 4.4).

There are also prospective non-randomised studies (quasi-experiments) with pre-test and post-test measurement in treatment and comparison groups using methods of analysis like statistical matching and/or double differences (DD). DD enables adjustment for time-invariant unobservable confounding at the level of the unit of analysis by design, and observable time varying confounding in adjusted analyses.[68] Thus, investigation of water supply and sanitation programmes, where hygiene messaging is often omitted, may credibly be done using double differences of individual or household panel data, where hygiene attitudes, an unobservable pre-existing confounder in analysis, may be considered fixed (time-invariant) and therefore controlled in analysis. In contrast, controlled before versus after studies, based on group level data, are not able to control for time-invariant sources of confounding at the individual level.

For example, the investigation of water connections in shantytowns on diarrhoea morbidity and water-related expenditures in Argentina used household fixed effects applied to survey data collected by the researchers at pre-test and post-test to estimate the double difference treatment effect (Galiani et al., 2009). Comparison neighbourhoods were chosen from among those who had applied to be connected but were not included for administrative reasons but were thought to be similar on observable characteristics. Other studies have used formal statistical matching methods like propensity score matching (PSM), to ensure units included have comparable pre-existing observable characteristics, also called common support (Heckman, 1998). For example, the prospective evaluation of a

[68] Difference studies can only adjust for unobservable confounding at the unit of analysis, hence it is important to distinguish studies where data analysis is at the individual or household level, from those where data analysis is conducted at the aggregate level such as the community, municipality or higher; studies based on aggregate level data are usually called controlled before-versus-after (CBA) studies.

community-driven development scheme in Maharashtra, India providing water and supply on costs, used PSM to match villages and difference-in-differences analysis to estimate the impact of the scheme on water-related costs (e.g., time to fetch water, time to use sanitation, medical expenses) (Pattanayak et al., 2010).

Some prospective studies have also been conducted with pre-test and post-test measurement in treated groups only (referred to as 'time series' study design by Cairncross et al., 1980). The most rigorous uncontrolled designs in theory use interrupted time-series, comparing trends before and after intervention (Shadish et al., 2002). Pre-test post-test designs that rely on a single data point rather than a trend are not usually considered credible. However, according to Victora et al. (2004), these approaches are valid where changes are measured a short period of time following the intervention, or the causal pathway is short, the expected effect is large, and confounding is unlikely. Where the causal pathway is longer, support for the relationship between intervention and outcomes can be made through examination of intermediate outcomes of mediator variable(s) along the causal pathway. This method is particularly powerful if the mediator variable(s) can be shown as unrelated to sources of confounding – that is, exogeneous, like an instrumental variable (Pearl and Mackenzie, 2018).

For example, a city-wide sanitation programme in Salvador, Brazil, laid over 2,000 km of sewer pipes, built 86 sewage pumping stations and connected 300,000 households to sewers between 1996 and 2004. Evaluation of the scheme used two time-series of children, one pre-intervention from December 1997 until April 1999 (which was the period before nearly all household sewer connections were made), and one post-intervention from October 2003 which was followed for eight months. Outcomes collected along the causal pathway included a hygiene practices index (Strina et al., 2006), intestinal parasite infections measured in stool samples (Barreto et al., 2010), household excreta disposal and open sewage nearby, and reported diarrhoea prevalence (Barreto et al., 2007). The study therefore combined interrupted time-series design with mediator analysis, as shown in hierarchical effect decomposition analysis (Genser et al., 2008; Bartram and Cairncross, 2010).

Other before-versus-after studies have been done retrospectively, using household recall to recover baseline data points. For example, evaluation of the St Lucia Poverty Reduction Fund, a CDD programme providing household water connections, measured time spent fetching water before and after (David, 2004). The validity of recall for time typically spent collecting water would invariably depend on the length of recall and the expectations operating due to self-reporting (see below Section 4.4.4).[69]

There have been parallel developments in methods of retrospective impact evaluations (of exposures and interventions) using observational data, including those that can address unobservable confounding based on knowledge about allocation rules that are external to participants (natural experiments). Examples include the following:

- Pure natural experiments in which treatment is assigned quasi-randomly by decision-makers using an exogenous mechanism. An example in WASH is the investigation in 1854 London of arbitrary exposure of households to sewage-contaminated water supply on cholera deaths (Snow, 1855). Morris et al. (2004) used quasi-random administrative errors in targeting to estimate the causal effect of the *Bolsa Alimentação* conditional cash transfer programme in Brazil on child linear growth.
- Regression discontinuity designs (RDDs) in which treatment is assigned by decision-makers based on a threshold on an ordinal or continuous variable (e.g., test score, age or date), and where 'as-if' random variation can be determined at the treatment threshold (Villar and Waddington, 2019). These are often undertaken retrospectively as natural experiments using observational data; for example, allocation of a village water supply programme in Guinea included an explicit rule that per capita costs should be less than Euro 100, which was used to estimate the impact on reported child diarrhoea for villages either side of the threshold using existing household survey data (e.g., Ziegelhöfer, 2012). In India, a Clean Village Prize with a substantial monetary incentive, was awarded to the leadership of Gram Panchayats achieving open defaecation free (ODF) status under the Total Sanitation Campaign. However, the value of the prize increased discontinuously according to population size, which

[69] Reporting is untransparent on this matter. The only information provided about recall is that "the evaluation was carried out very soon after the completion of the sub-projects" (David, 2004, p.ix) which built or extended water systems which had been in operation for between three and 30 months. So, in this instance, the minimum recall appears to be three months and maximum 2.5 years.

Spears (2013) used in estimating the effect of the incentive on ODF status, stunting rates and infant mortality. RDDs do not have to be designed retrospectively, but they usually are, in part due to the large samples needed for statistical precision (Goldberger, 1972). In addition, the administrative errors and discontinuities which they exploit have rarely been implemented with a view to facilitating evaluation. Rather, they are discovered and used opportunistically by the evaluators.

- Instrumental variables (IV) estimation in which investigators identify 'as-if' randomly distributed exogenous factors which are correlated with treatment assignment but do not determine the outcome of interest, except through treatment (e.g., Greenland, 2000). IV estimation, and related approaches,[70] is often done of exposures, although it is also used in intervention studies including RCTs. IV estimation uses multiple-stage regression modelling (e.g., two-stage least squares, 2SLS) or simultaneous equations maximum likelihood (e.g., bivariate probit). Exogenous variables used in IV estimation include the variables mentioned above, such as randomised assignment or encouragement, where instrumental variables estimation is used to account for non-compliance (Imbens and Angrist, 1994).[71] Other studies have used random variation in weather or climate conditions to estimate the impact of diarrhoea and dehydration in childhood on hypertension, a major cause of heart disease and death in adulthood (Lawlor et al., 2006), and topography to estimate the impact of dams on increasing poverty in India (Duflo and Pande, 2008), although the validity of topography in satisfying the exclusion restriction has been questioned (Deaton, 2010).[72]

[70] For example, structural nested modelling (e.g., Brumback et al., 2014) and switching regression models (e.g., Lockshin and Sajaia, 2004).

[71] The relationship of interest in encouragement studies is not usually the pragmatic question about the effect of such encouragement, but rather the mechanistic question about the effect of the intervention in people who are responsive to encouragement. A randomised encouragement study can be analysed conventionally (using intention-to-treat) or using instrumental variables estimation.

[72] Geographical factors such as distance are often used (e.g., Newhouse and McClellan, 1998). However, location is endogenous – at least in the long-term, people are able to move to gain access to services, and location itself may explain differences in outcomes such as for geographically marginalised groups; i.e., the 'exclusion restriction' is not usually satisfied. Hence distance of participant to facility is often not a valid instrument. McKenzie et al. (2010) used distance to application centre in Tonga as a valid instrument for immigration to New Zealand, arguing that while it affected participation in the lottery enabling emigration to New Zealand, it did not affect the counterfactual outcome (income), at least for those that lived on the small main island of Tonga. See Chapter 5 Section 5.3.4.

- Interrupted time series (ITS) where the trend in outcomes data is measured pre- and post-intervention (e.g., Duflo et al., 2015). These studies are often done on group level data, although analyses on individual data would increase statistical power. They are considered particularly credible when contemporaneous data are available on a control group (Shadish et al., 2002).
- Double differences estimation[73] applied to longitudinal panel data – or pseudo-panel (repeated cross-section data) under particular conditions (Verbeek, 2008) – of outcomes collected at pre-test and post-test in treatment and comparison. These studies are usually done at individual level and may be combined with statistical matching of participation at group level to determine the comparison group sample. However, an example of a group-level panel study using observational data is the investigation of the effect of water privatisation in on child mortality rates in municipalities in Argentina (e.g., Galiani et al., 2007). An example of analysis of a pseudo-panel using observational data at individual level (with individuals matched using PSM), is the study of diarrhoeal mortality due to urban water supply and sewerage improvements in Ecuador (Galdo and Briceño, 2005).

Observational studies with selection on observables evaluate outcomes in the presence and absence of treatment using parametric methods like OLS regression analysis. Where statistical matching (e.g., propensity score matching, PSM) is used to compare treated and untreated observations on observable characteristics, outcomes are compared non-parametrically. Frequently, these studies aim to estimate the effects of an exposure rather than an intervention. They may be applied to cross-section data such as the evaluation of the impact of piped water supply on child diarrhoea using DHS in India (Jalan and Ravallion, 2001), Egypt (Roushdy et al., 2011) and the Philippines (Tan et al., 2012), or case control data retrospectively compiled, as in investigation of latrine access on diarrhoea (Daniels et al., 1990a). An identifying characteristic of matching is that it is done on observable factors collected at baseline, or time-invariant factors measured at endline, which can be credibly argued as strongly correlated with unobservable sources of confounding – that is, the assumption of 'unconfoundedness', also called 'strong ignorability' (Imbens and Wooldridge, 2009). However, single

[73] This category implicitly includes approaches like 'triple differences' and fixed- or random-effects analysis of individual level longitudinal panel data.

difference estimation applied to case-control, cohort, or cross-sectional data (or in PSM when matching is on baseline characteristics) is not able in theory to control for time-varying or time-invariant unobservables, and as indicated below in Section 4.4.2, there may be important unobservable sources of selection bias.

4.4 Internal validity in impact evaluations

The ability of impact evaluations to produce valid causal inferences depends on both study design, which in turn depends on underlying assumptions (which may be untestable for some designs, especially NRS), and quality of implementation of the study, which is verifiable largely based on reporting (Littell et al., 2008). High quality systematic reviews set explicit study design inclusion criteria, and then transparently appraise included studies based on the quality in which they are designed and implemented (internal validity) (Higgins and Green, 2011; Waddington et al., 2012). Some reviews also assess external validity, or the relevance of the evidence (e.g., Waddington et al., 2009), covered in the next section.

Study designs with selection on unobservables, like RCTs, natural experiments and RDDs, are usually considered more credible at identifying causal relationships (internal validity) in theory, than studies which assume unconfoundedness (Shadish et al., 2002; Imbens and Wooldridge, 2009; Dunning, 2012). The main complications of many non-randomised studies, however, are the untestable assumptions and the need for diagnostic and falsification analyses. This makes them "more susceptible to influence from researcher expectations and hypotheses that can bias study results towards what is expected or desired rather than what is true" (Chaplin et al., 2018, p.7).

To understand the assumptions underlying impact evaluation methods, it is important to distinguish the treatment effect estimate that is being sought. The assumptions underlying validity of effect of treatment assignment, or intention-to-treat (ITT), are different from those underlying the effect of starting and adhering to treatment (per-protocol effect), or treatment-on-the-treated (also called average treatment effect on the treated, ATET, complier average causal effect, CACE, or local average treatment effect, LATE) (Sterne et al., 2016; Swanson et al., 2017). For instance, let Z be a

variable determining assignment, *T* a variable representing treatment status, and *O* the outcome of interest for observations (*i, j…n*). There are three overarching assumptions underlying the internal validity of RCTs and 'as-if randomised' studies (natural experiments, RDDs, instrumental variables, IVs) to estimate the effect of treatment ('per-protocol' effect):

1) Relevance or fixed (predictable) relationship between *Z* and *T*: in the case of RCTs and natural experiments (and when RDD and IV are used to estimate the 'global' average treatment effect, ATE), there is a homogenous relationship between *Z* and *T* across all units; for RDDs used to estimate the local average treatment effect (LATE) and instrumental variables used to estimate the complier average causal effect (CACE), there is a nonzero and monotonic causal relationship between *Z* and *T* ('no defiers' or the absence of 'no-shows' and 'crossovers') (Bound et al., 1995).
2) Independence of observations (*i, j*) (or stable unit treatment value assumption, SUTVA) (Chiba, 2010): *Z* for treatment unit *i* does not affect *T* for treatment unit *j* (no subversion of the assignment process or selection bias into the study); *T* for treatment unit *i* does not affect *Y* for treatment unit *j* (absence of 'spillover effects'); and there is no variation in *T* across observations (e.g., due to problems in implementing the intervention of interest, differential attrition, time-varying non-adherence in sustained interventions, or measurement error).
3) Externality and exogeneity of *Z*: that is, *Z* is external to *Y* (it is not affected by *Y* or any of its causes) and only affects *Y* through *T* (the 'exclusion restriction').[74]

Typically, appropriate instruments are usually generated through natural experiments or random assignment of the treatment in the case of RCTs with imperfect compliance. In the absence of these conditions, it is difficult for validity to be verified. For example, the internal validity of instrumental variables estimation rests on three main conditions. Firstly, the instrument must be relevant. It must significantly affect participation in the programme. The greater the correlation between instrument and participation, the more accurate the estimation. Secondly, SUTVA must be satisfied, which is usually done for IV by assuming a predictable (monotonic) rather than fixed relationship between instrumental variable and treatment status. Thirdly,

[74] The degree of homogeneity of the relationship between *T* and *Y* across individuals induced to treatment by *Z* is also of interest for external validity (Angrist and Pischke, 2009).

the instrument must be exogenous: that is, it is external, meaning the absence of a simultaneous or reverse causal relationship between the instrument and the dependent variable; and it must only affect the outcome through participation (the exclusion restriction) (Deaton, 2009). For example, river gradient has been used as an instrument for the effect of construction of dams on agricultural production and poverty rates in India (Duflo and Pande, 2007) and land gradient used as an instrument for access to water plants providing improved water quality on diarrhoea and nutrition (Zhang, 2011). River and land gradient are not affected by the variables being explained and are clearly external. However, the sufficient condition to satisfy exogeneity is that they should not affect outcomes directly, or through another route than the intervention. Gradient may theoretically affect the outcome in both cases – in the case of agricultural production and poverty by presenting difficulties to farmers living on marginal uplands; and in the case of health outcomes by presenting difficulties to obtain sufficient water supply.

An example of breach of the exclusion restriction in a trial would be the effect of participant expectations on behaviour (e.g., the Hawthorne effect) or reporting (e.g., social desirability bias in open trials) which may lead to the estimation of an effect, even in the absence of an efficacious intervention of interest. There is also some debate in the literature about whether it is necessary to differentiate intervention effects from pure placebo effects. Arguably, in social interventions requiring behaviour change from participants, expectations may form an important mechanistic component in the process of behaviour change, determining uptake and adherence. Therefore, isolating expectation effects (such as placebo effects) from other causal mechanisms may be less relevant (Waddington et al., 2012). However, factors relating to motivation of those being observed regarding behaviour or reporting are still of major concern in trials.[75]

Double differences estimation can control for time-invariant confounding only, at the unit of analysis. Suppose data are collected in two periods for participants and non-participants, at pre-test time $t=0$ and post-test $t=1$:

[75] See also Anand et al. (2020).

$$Y_i^{t=0} = \beta X^{t=0} + \gamma T_i^{t=0} + \mu_i + \varepsilon_i^{t=0} \qquad (4.14)$$

$$Y_i^{t=1} = \beta X_i^{t=1} + \gamma T_i^{t=1} + \mu_i + \varepsilon_i^{t=1} \qquad (4.15)$$

where Y_i^t is the outcome of interest in each period for participant i, X_i^t is a set of measured covariates, and $T_i^{t=1}$ is participation in the programme (equal to 0 for participants and non-participants at pre-intervention time t=0, 1 for post-intervention participants at t=1, and 0 for non-participants at t=1), μ_i is a variable capturing time-invariant unobservable characteristics for each participant i, and ε_i^t is the error term. Subtracting the former from the latter gives:

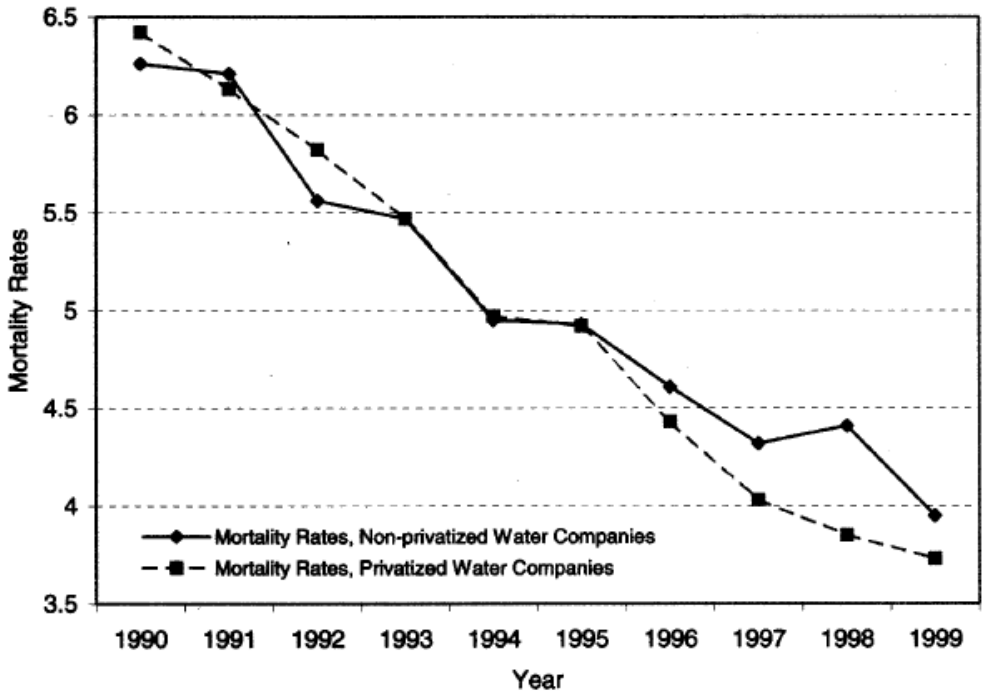
$$\begin{aligned} Y_i^{t=1} - Y_i^{t=0} &= \beta\left(X_i^{t=1} - X_i^{t=0}\right) + \gamma\left(T_i^{t=1} - T_i^{t=0}\right) + \left(\mu_i - \mu_i\right) + \left(\varepsilon_i^{t=1} - \varepsilon_i^{t=0}\right) \\ &= \Delta Y_i = \beta \Delta X_i + \gamma T_i^{t=1} + \Delta \varepsilon_i \qquad (4.16) \end{aligned}$$

Equation (3.16) can be estimated using ordinary least squares (OLS), or it can be rearranged to include baseline outcomes as an independent variable and estimated using fixed effects panel data regression. As shown, this allows control for unobservable time-invariant factors at the unit of analysis, since $\mu_i - \mu_i$ cancels out in equation (4.16). Therefore, provided units are measured at the household or individual level, some sources of confounding – such as household hygiene or water consumption in an evaluation of latrine provision – will be 'differenced out'. However, the method is susceptible to bias when participation and outcomes are jointly explained by an unobservable time-varying characteristic (including attrition), as well as any unobservables at further disaggregated levels of analysis (e.g., unobservables at the individual level if analysis is at group level). Unobservable confounding at individual level is likely to be more important when self-selection is an important determinant of treatment, for example due to household income or hygiene behaviours. In contrast, where investigation is of the impact of interventions placed by programme planners, such as community level extension of water supply (e.g., installation of handpumps) or sanitation infrastructure (e.g., sewer connections), household income or hygiene behaviours are likely to be less important determinants of participation. In addition, a shorter time frame might effectively fix factors such as income.

As with other NRS, the approach therefore needs to incorporate falsification methods, such as 'placebo interventions' or 'placebo outcomes'. A common approach, which could also be called 'placebo time periods', compares the evolution of outcomes among treated and untreated units before intervention. This can be reviewed via visual inspection or formally tested using a 'leads and lags' approach (Autor, 2003). If outcomes are perceived to have equal secular trends during periods prior to intervention, and trends diverge post intervention, this is suggestive of an intervention effect.

For example, Galiani et al. (2005) evaluated the impact of privatisation of water supply in municipalities in Argentina on child mortality, following an increase in the rate of privatisation of water supplies by local governments in 1995 following re-election of the central government. The authors present (and verify using statistical analysis) equal secular trends in mortality reduction between 1990 and 1995, following which the rate of reduction in mortality rates increased in poorer municipalities with privatised water supply (Figure 4.2).

Figure 4.2 Evolution of mortality in municipalities in Argentina



Source: Galiani et al. (2005).

In addition, in analysis of a 'placebo outcome', the authors examine cause-specific mortality, finding that neonatal and infectious diseases fell in treated municipalities, but not accidents, cardiovascular disease or cancer, which is

consistent with the reduction in water-related disease transmission. The authors argue (and present evidence) that the reduction in mortality is due to increased investment by the private water providers, which improved access to water among poorer income groups.

Other reviews have argued why it is not just parallel trends that need to be established at baseline but also levels. For example, Schmidt (2017) presents several alternatives for the evolution of outcomes, suggesting that DD will be more reliable with a statistically matched sample, including matching on baseline outcome.

In systematic reviews examining questions about the effects of interventions, assessment of internal validity is done in risk-of-bias assessment. Risk-of-bias tools provide the criteria to enable reviewers to evaluate transparently the likelihood of bias, for particular bias domains (e.g., confounding, selection bias, performance bias, bias in data collection and reporting biases). The diarrhoeal disease measurement literature has long identified factors such as confounding, recall bias and failure to collect intermediate outcomes as important sources of bias when diarrhoea is measured by self-reporting (Blum and Feachem, 1983). More recent literature has articulated sources of bias which are common in RCTs and NRS (e.g., Sterne et al., 2016).

Table 4.6 compiles common sources of bias in WASH impact evaluations, which, in the broadest sense, are all sources of confounding in accurately measuring the causal relationship between intervention and outcome. They are grouped into four main categories or domains of bias affecting internal validity: confounding and selection bias (confounding in study design and implementation); performance bias or bias due to departures from intended interventions (confounding in programme implementation); bias in measurement of intervention or outcomes (confounding due to measurement error); and selective analysis and reporting (confounding due to publication bias). A final domain, adequacy of the sample size, affects accuracy of statistical testing in small samples of interventions assigned to dependent observations (e.g., at village level).[76] These biases are discussed in Sections 4.4.1 to 4.4.6.

[76] As noted in Table 4.6, this may also affect bias due to baseline confounding, where an insufficient sample size leads to groups which are unbalanced on pre-existing criteria.

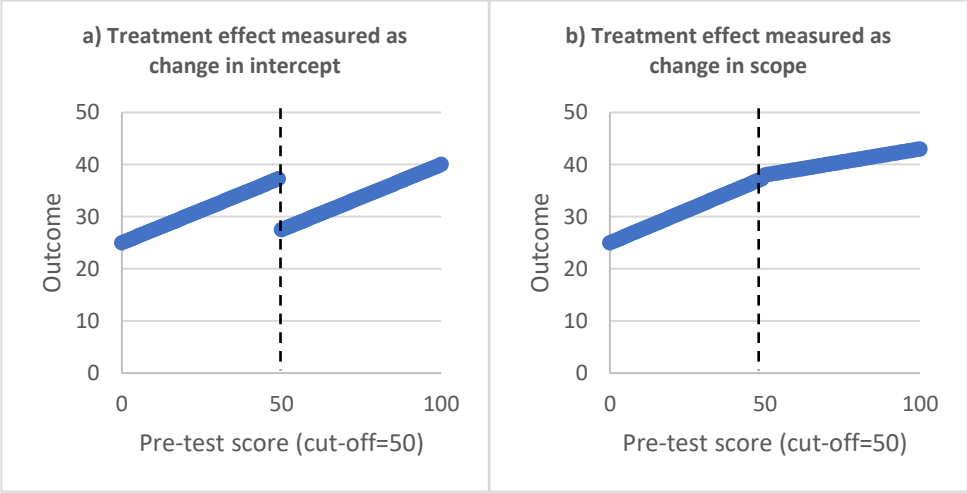
It is worth noting here that, although the underlying domains of bias (confounding, selection bias, departures from intended interventions, measurement error, and selective analysis and reporting) are relevant across all designs, whether randomised or non-randomised, prospective or retrospective, the criteria used to verify them will differ. In particular, the 'signalling questions' on which each of these propositions are verified will depend on the assumptions underlying each approach. For example, there is renewed interest in the use of RDD, also called regression discontinuity or 'cut-off-based design' (Shadish et al., 2002), as a method of programme evaluation, including in WASH. A local government authority may set a threshold on per capita unit cost estimates to determine whether village water connections are cost-effective (Ziegelhöfer, 2012). While villages at the extreme ends of the population size distribution are likely to be very different (e.g., small size reflecting remoteness), villages on either side of the cut-off threshold should be very like one another. Comparison for this subset of villages may therefore be made and any treatment effect shown as a discontinuity (or break) in outcomes between treated and untreated groups at the point of intervention.

In RDD, treatment is assigned *ex ante* according to a known rule – specifically, a threshold on a scale variable measured among participating units at pre-test. Units scoring on one side of the threshold subsequently receive treatment, while those on the other do not. The treatment effect is estimated by comparing observations from different units observed contemporaneously, immediately on either side of the threshold. Different types of assignment variables have been used in RDD analyses (Hahn et al., 2001; Dunning, 2012; Moscoe et al., 2015) such as test scores (e.g., continuous biomarkers in medicine), programme eligibility criteria (e.g., poverty index), age (e.g., birth date), size (e.g., hospital or school size), and time (e.g., date of a policy or practice change). In geographical discontinuity design (GDD), exposure to the treatment depends on the position of observations with respect to an administrative or territorial boundary (e.g., Galiani et al., 2017).

In the basic design, assignment to treatment and comparison is based on the observational unit's pre-test score on the continuum, relative to the assignment threshold (Bor et al., 2014). Figure 4.3 presents two simple

examples of the relationships between an assignment variable (pre-test score with cut-off set at 50) and outcomes. Sometimes, it is the researcher who designs the study prospectively (Buddelmeyer and Skoufias, 2004). However, discontinuity assignment is usually exploited in natural experiments because of natural processes of policy and practice. For RDD to produce internally valid estimates, the minimum criterion is 'exchangeability at the threshold' (Bor et al., 2014) – that is, the potential outcomes would be the same on average if treated units had been untreated and untreated individuals had been treated, as would be the case in a well-conducted RCT. One common way that this is violated is if the assignment variable itself is precisely manipulable by participants or implementers, at least over the sub-sample of observations around the cut-off threshold. Threats to validity may arise where there is public knowledge among programme participants of a manipulable assignment variable, or where practitioners are able to assign to treatment on a discretionary basis.

Figure 4.3 Examples of RDD



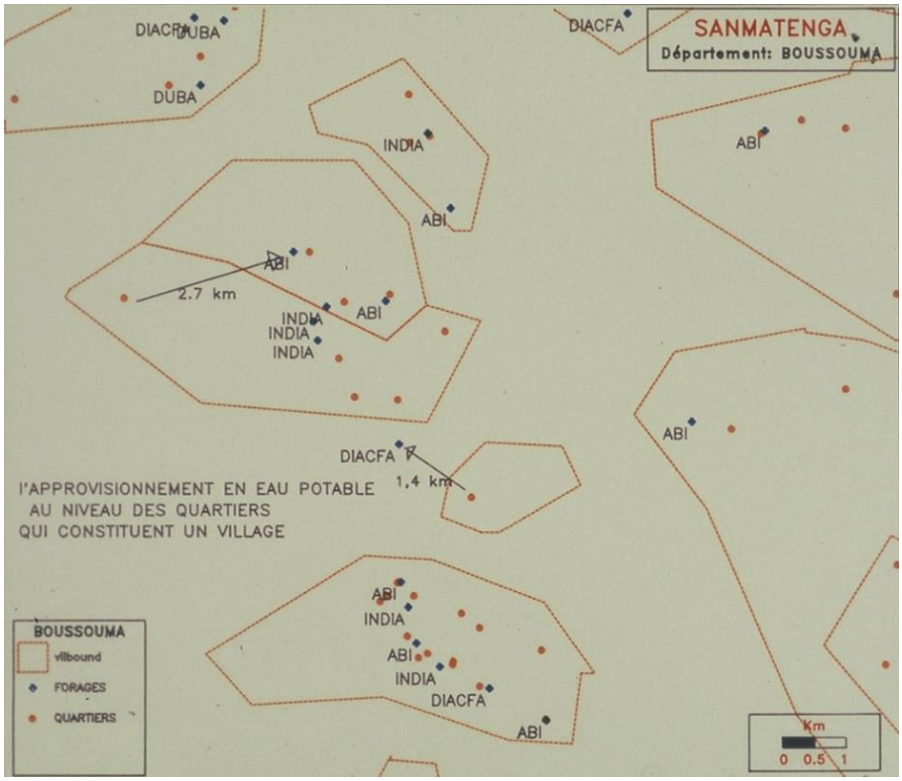
Source: Villar and Waddington (2019).

Examples of thresholds on continuous variables that one might think could potentially be exploited in geographical discontinuity design are where decisions to use a water or sanitation intervention at household level are allocated around a 'source-choice boundary' (Cairncross et al., 1980). Households immediately on either side of the source-choice boundary are expected to sort to collect water from either source, some choosing either source on each side. Were this to form a natural experiment of the impacts of source water, one would have to assume some random variation in sorting at the boundary.

Another case would be where treatment decisions are taken centrally according to a threshold. A common threshold that might be thought exploitable in RDD analysis is the use of below poverty line (BPL) cards to allocate subsidised access to latrines under the Total Sanitation Campaign in India (Dickinson et al., 2015). However, BPL status is used to allocate other benefits, hence it cannot be used to identify the effect of latrine access on outcomes that might be affected by other non-WASH interventions (although it could be used to identify the effect of a range of interventions on these outcomes).

An example, of handpumps and Guinea worms in southern *Sanmatenga* province in Burkina Faso in the early 1990s, is illustrative. In Figure 4.4, each polygon shows the area of a village, orange dots are hamlets and blue dots boreholes with handpumps. In some hamlets, people needed to travel several kilometres, or into the next village, to find their nearest handpump. In the rainy season, stagnant water would collect in ponds which people would use to obtain water as they were closer to home, from where they were likely to spread Guinea worm disease (*dracunculiasis*) by walking in water with a worm exposed on the leg or contract it by drinking infected water.

Figure 4.4 Water supply in villages in Burkina Faso



Source: Sandy Cairncross, pers. comm.

New handpumps were more likely to be installed by the Government in villages where more numerous worm infections had been reported through the community-based surveillance system.[77] If *dracunculiasis* incidence had been testable objectively (e.g., through health facility reports), and there was a direct correlation between incidence at pre-test and intervention, it would have been possible in theory to measure the effect of handpumps on disease as a discontinuity in the relationship between pre-test and post-test. However, village chiefs, who knew about the assignment rule, were incentivised to overreport incidence to obtain resources for handpumps.

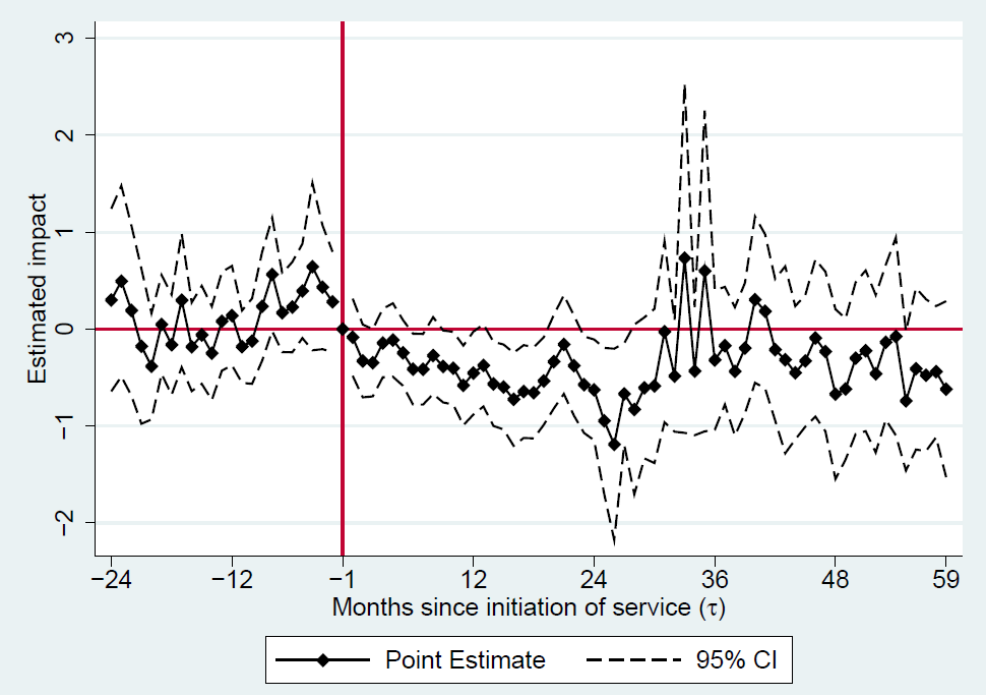
Assessing whether the forcing variable measured at pre-test is manipulable is therefore equivalent to assessing subversion of randomisation when random allocation is not concealed until after recruitment in RCTs and cluster-RCTs. Hence, participants should either be blinded to the value of their assignment variable or unable to manipulate it (and practitioners should not be involved in assignment, or unable to manipulate it). Assignment variables which participants have manipulated include reported income, which may be incorrectly reported or manipulated to gain eligibility to programmes (Buddelmeyer and Skoufias, 2004). Establishing non-manipulation is easiest for the sample of observations closest to the threshold where there is random error in measurement of the assignment variable (Goldberger, 1972). One test typically used to check manipulation is due to McCrary (2006), which examines discontinuities in the density of the forcing variable at the cut-off.

As with other NRS, confirmation tests (e.g., comparison of covariate means either side of the threshold) and falsification methods are an important component of internal validity assessment. Recruitment into the study of a 'pure control' group that is subjected to the same informed consent and data collection also has the advantage of enabling measurement of motivation biases in a prospective RDD. The addition of falsification methods such as a non-equivalent dependent variable function ('placebo outcome') can also be included, as well as tests for 'placebo discontinuities' at different thresholds of the assignment variable, which can help rule out the existence of a chance relationship.

[77] See Cairncross et al. (1996) for an overview of community participation in Guinea worm eradication programmes.

A similar design to RDD is the interrupted time-series (ITS). Figure 4.5 shows a reduction in diarrhoeal infections requiring medical assistance in the period following installation of piped water supplies and latrines to all households in NGO Gram Vikas (GV) villages in India. The design incorporates good principles of ITS, including more than six periods of outcomes data collection pre- and post-test (Freitheim et al., 2015) and an observable effect of the RDD and ITS immediately after water connections were turned on.[78] ITS and RDD are sometimes seen as equivalent approaches, especially in the case of regression discontinuity in time (RDiT), where the assignment variable is time (Hausman and Rapson, 2018).

Figure 4.5 Cases of diarrhoea treated monthly in Gram Vikas villages



Note: impact variable is normalised at the month (-1) immediately prior to installation of water supply; all estimates are relative to month (-1).

Source: Duflo et al. (2015).

[78] The outcomes data were collected by GV programme staff, for standard monitoring – not for the purpose of an evaluation – and personnel were sanctioned for misreporting. The data were found accidentally: "the paper forms were locked in a closet when they were uncovered by the research team during a visit to discuss an unrelated evaluation" (Duflo et al., 2015, p.13).

Table 4.6 Methodological problems affecting internal validity in WASH impact evaluations

	Type of bias	*Explanation*	*Example*
Confounding and selection bias (study design and implementation)	Absence of control (baseline confounding)*	Absence of a control is problematic, or where control is not comparable due to pre-existing differences (baseline confounding).	Control (comparison in NRS) is required to adjust for confounding, for outcomes that do not occur immediately (e.g., where access or use of technology in response to intervention is delayed) or where the causal pathway is long (e.g., health and socioeconomic outcomes, and most behavioural outcomes, with the exception of time use).
	Selection bias^	Some eligible treatment units or follow-up periods are excluded from data collection or analysis.	Selection of treatment units or follow-up periods causes bias when exclusion of units is correlated with outcome. Selection bias out of the study (attrition) is problematic in longitudinal studies including trials. Selection bias into the study is problematic in studies designed retrospectively (after implementation of intervention).
	Confounding (time-varying)*	Inadequate control for confounding (time-varying confounding); the confounders will vary depending on the intervention and outcome of interest.	A multi-country observational study of water and sanitation and reported diarrhoea (Esrey, 1996) excluded water supply functioning and use, hygiene practices and socioeconomic status, impairing the causal inferences made (Cairncross and Kolsky, 1997). Time-varying confounding is less problematic in evaluations of 'baseline interventions', interventions that are implemented at one point in time at the start of the study (e.g., deworming). However, time-varying confounding is particularly important when trying to estimate the per-protocol effect in evaluations of 'sustained interventions' that require continued adherence to treatment, including in RCTs.
	Failure to analyse by age*	Age-specific analysis is necessary.	Outcomes, particularly diseases like diarrhoea, are unevenly distributed among age groups. For example, diarrhoea is usually most incident in young children. Behaviour and facility use also depend on age, sex, disability and cultural factors.
	Failure to account for seasonality*	Outcomes vary by season, especially diarrhoeal diseases and parasitic worm infections.	Measurement should take place during the same period in treatment and control, to avoid confounding by seasonality, and preferably during the season of peak incidence for the outcome, where the effect size of the intervention will be greatest and hence most detectable.

	Type of bias	*Explanation*	*Example*
Departures from intended interventions	Performance bias^	No-shows, crossovers, spillovers, and implementation fidelity.	No-shows and crossovers (contamination), together called switches, occur where individuals receive a treatment different from that assigned. Assessment should therefore be made of the extent to which these are accounted for in design or analysis, such as through ITT estimation of the effect of assignment to treatment, or instrumental variables estimation to measure the per-protocol treatment effect (CACE). Spillovers occur when members of the comparison group are exposed to treatment indirectly, through contact with treated individuals; spillovers are potentially problematic in all controlled studies measuring communicable disease. Cluster-level analysis may be required to ameliorate these sources of bias and an assessment of the geographical or social separation of groups needed. Fidelity of implementation to treatment protocols, may also affect exposure of study participants to the intervention, and therefore outcomes. This source of bias has also been called Type III errors (Dobson and Cook, 1980).
	Motivation bias	Hawthorne, John Henry and survey effects.	Hawthorne and John Henry effects alter the motivation of participants who are aware they are part of a trial. 'Survey effects' may operate whereby groups are sensitised to information that affects outcomes through survey questions and then subjected to repeated measurement (Zwane et al., 2011). They are less likely to affect motivation where data are collected outside of a trial situation with a clear link to an 'intervention', and unlikely to be relevant when data are collected at one period of time as in a retrospective cross-sectional only.
Measurement error for interventions	Bias in measurement of intervention^	Intervention recall may be problematic, especially where information on dose, frequency, intensity or timing are needed.	This is not usually considered problematic where information is collected at the time of the intervention from sources not affected by outcomes (e.g., enumerators). It is problematic where information about treatment status is obtained after implementation from participants or practitioners who may misremember in recall or have an incentive to misreport.
	Failure to record facility usage*	Access is the necessary condition but usage is the sufficient condition to improve outcomes.	It is important to measure adherence in sustained interventions requiring continued behaviour change. Systematically obtained observational data on usage is preferred to reported data (e.g., quantity of water used, latrine facility use by children).

	Type of bias	*Explanation*	*Example*
Measurement error for outcomes	Bias in measurement of outcomes due to lack of blinding and/or participant reactivity^	In open trials, where individuals are aware of their treatment status, lack of blinding of participants may lead to bias in reported outcomes measurement.	Lack of blinding is usually only problematic in trials where outcomes data are reported, and where participants can clearly identify an intervention due to informed consent (Schmidt, 2014). In longitudinal studies of sustained interventions, participant fatigue may cause unwillingness to engage further with survey enumerators (outcome assessors) (the "Bugger-Off Effect") (Clasen, 2013; see also Schmidt and Cairncross, 2009). Where enumerators are not blinded to intervention, they may induce desired reporting by participants (the "Clever Hans Effect") (Beath et al., 2013). Desirable reporting has also been found for easily modifiable behaviours (clean hands and presence of soap at handwashing station) when outcomes are observed due to participant 'reactivity' in longitudinal survey, in the absence of hygiene interventions (Arnold et al., 2015). Double blinding of participants and outcome assessors to intervention is usually impossible, with the exceptions of anti-bacterial hygiene interventions (Larson et al., 2004) and some household water treatment devices, although even this may be difficult due to water turbidity (Boisson et al., 2010). Blinding of outcome assessors may be possible, where controls are provided a placebo 'intervention' that does not affect outcomes of interest, e.g., children's books, notebooks, pens and pencils in a household water treatment and hygiene trial in Pakistan (Luby et al., 2004). However, blinding of others involved in the study with reporting incentives such as data analysts may be more feasible. For example, a study in Brazil blinded data analysts to intervention status in laboratory measurement of 20 percent of stool samples (Moraes et al., 2004).
	Health indicator recall*	Recall is hampered by knowledge of the person providing information about others and their memory of events. Self-reporting is hampered by expectations.	Recall of others' experiences is more likely to be accurate if done by a child's carer. A recall period for diarrhoeal morbidity exceeding two weeks is considered unreliable, and it should preferably be no longer than 48 hours. Expectations include over-reporting of 'desirable' behaviours linked to treatment (Manun'Ebo et al., 1997), the promise of treatment in control groups or, in either group, underreporting due to shame or unwillingness in health studies to submit blood or stool samples.
	Health indicator definition*	Indicators need to be defined precisely, to ensure that they measure the same construct across individuals.	For example, diarrhoea may be defined clearly to study participants as three or more loose or watery stools, with or without blood, in 24-hours. Consistency may also facilitate cross-study comparisons of outcomes. Measurement bias may also occur if data collection instruments are not the same between treatment and control or over time.

	Type of bias	*Explanation*	*Example*
Selective analysis and reporting	Bias in selection of the reported result^	Selective reporting of outcomes (e.g., among multiple possible outcomes collected), selective reporting of results from sub-groups of participants (e.g., among multiple participant groups), or selective reporting of methods of analysis (e.g., multiple estimation strategies or specifications).	Selective reporting is particularly likely to be prevalent in retrospective evaluations based on observational datasets (e.g., with many IV analyses), but may also arise in prospective studies where the method of analysis, outcomes or sub-groups are chosen based on results (e.g., Freeman et al., 2014). Presence of a study protocol (pre-analysis plan) can help determine the likelihood of bias, although it is recognised that many such studies still do not contain such plans, particularly non-randomised studies and natural experiments, nor is it always possible to fully specify all models in advance.
Bias in statistical analysis	Adequacy of sample size (one-to-one comparison)*	Use of a small number of treatment units without control for dependency within units (e.g., one village in each treatment group).	Impact evaluations need sufficient independent observations to ensure covariate balance (and estimate effects with statistical precision). Where interventions are delivered at cluster level, and especially where transmissible disease is measured, observations within clusters are likely to be dependent. Information on the intra-cluster correlation coefficient is also needed to estimate the effective sample size in prospective studies and conduct statistical tests. It is also worth noting that small effective sample sizes also affect the likelihood of achieving balance (e.g., Katz et al., 1993), and therefore may introduce confounding (White, 2013).

Notes: * from Blum and Feachem (1983); ^ from Sterne et al. (2016).

However, ITS and RDiT use different treated and untreated samples, which makes a difference to the length of follow-up period over which treatment effects can be credibly estimated. In ITS, the same participating units are followed up over time, and the treatment effect is identified through variation in exposure to treatment over time, sometimes with respect to an untreated comparison (Shadish et al., 2002; Somers et al., 2013). ITS is most credible in estimating treatment effects for observations immediately after the time of intervention in comparison with their values immediately before (i.e., the short-term effect). In contrast, in RDiT, the treatment effect is estimated by comparing observations from different units measured at the same time (or follow-up period); the comparison is made up of units who were eligible immediately before or after a threshold date on which a policy or practice change occurs. However, the outcome for those units could be assessed many years later.

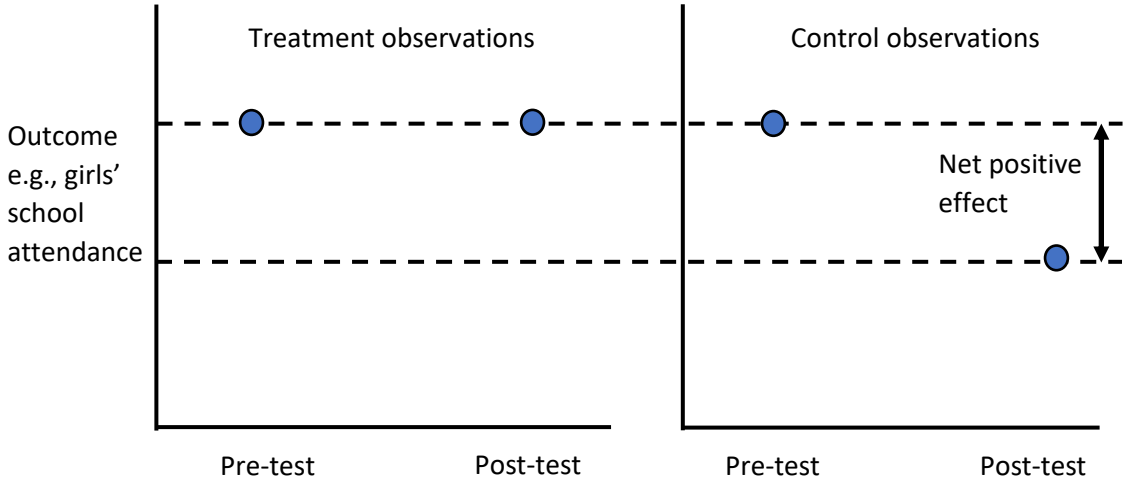
4.4.1 *Confounding*

Confounding bias occurs when factors which predict the outcome also determine receipt of intervention. This includes self-selection to intervention by participants (e.g., based on need) or practitioners (e.g., based on eligibility), or programme placement decisions by planners (e.g., on the basis of geographical unit). Confounding bias is nearly always thought more problematic in non-randomised and observational studies than in trials (e.g., Sterne et al., 2016). For example, confounding is likely to be problematic in retrospective studies where baseline data cannot be collected to ensure balance on pre-existing covariates. As noted by Blum and Feachem (1983, p.360): "[e]ven if no health improvements are detected, no conclusions can be drawn because it might be that health would have deteriorated without the water or sanitation investment or, conversely, that health would have improved and the water supply or excreta disposal facilities increased transmission of certain infections." For example, a pre-test post-test evaluation is not able to detect any change in attendance before and after installation of separate latrines for girls in schools. However, during the intervention period, the village water pumps had broken down on some days due to a drought, so girls needed to help fetch water from other sources further away on those days. The net effect of the latrine installation was to protect adolescent girls from staying at home during their period days, but not on days when they needed to fetch water. Therefore, only when

comparing against individuals in control villages, also affected by the drought, can the protective net effect of the scheme be observed (Figure 4.6).

It is often thought important that controls selected are identical to treated observations (samples are balanced on observables and unobservable characteristics).[79] In RCTs, ensuring balance between treatment and control also means randomisation over a sufficiently large sample to ensure pre-existing characteristics are equal on average; in effect, both treated and untreated observations are taken randomly from the same underlying population. In NRS, which by definition involve non-randomly selected treatment observations, the comparisons must therefore also be selected non-randomly, for example by matching on pre-existing observable characteristics.

Figure 4.6 Use of a control group to measure the net effect



Statistical matching is used alone or in conjunction with other prospective or retrospective designs, including RCTs where pair-wise matching of observations may be done before randomisation to improve efficiency in small samples (e.g., King et al., 2009; Nicholson et al., 2014), and double differences (as noted above). Propensity score matching (PSM) is an efficient matching estimator that compares units based on predicted scores on a participation equation constructed of observable characteristics (Rosenbaum and Rubin, 1983). Formally, the impact estimator, the average treatment effect on the treated (ATET) is calculated for unit *i* as the difference

[79] This is not necessarily the case for DD, IV and RDD, where statistical methods are used to obtain balance (see discussion below).

in the expected value of outcome *Y* if the unit participated in the programme *T=1* and if they did not *T=0*:

$$ATET_i = \left(\bar{Y}_i \mid T = 1\right) - \left(\bar{Y}_i \mid T = 0\right) \qquad (4.17)$$

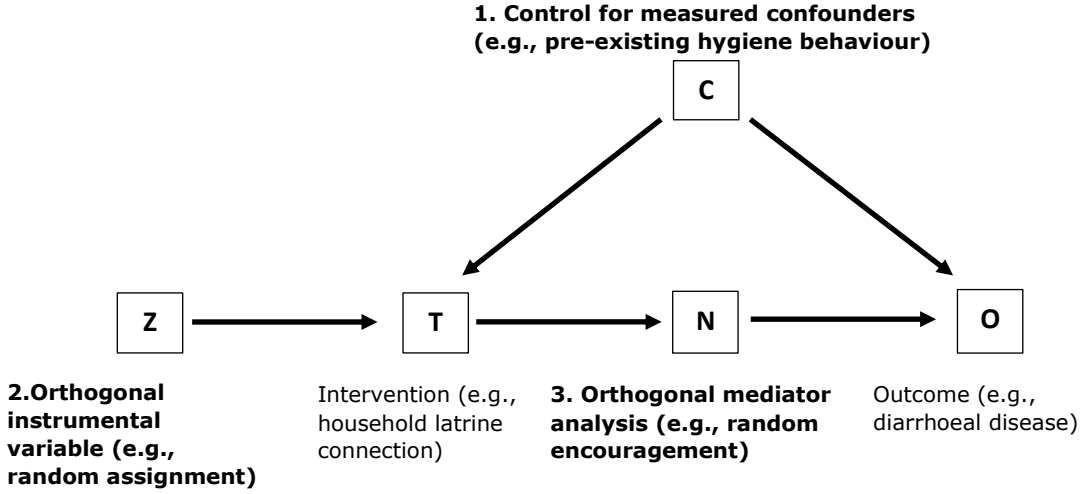
The "fundamental problem of causal attribution" (Holland, 1986; cited in Rubin, 1990, p.478) is that it is not possible to observe *Y* for the same individual at the same time. To overcome this problem, PSM identifies for participants a group of comparison units with the same probability of participating based on observable characteristics, but who do not participate in practice. There are three steps in the approach. In the first step, the propensity score equation is estimated by regressing the probability of participation on a set of observable characteristics. This is often done in logistic regression, but probit and survival models are also used. In the second step, the predicted probability of participation for each unit is obtained from the coefficients in the participation equation, which is used to match treated and comparison observations. Various propensity score matching techniques are used, including 'nearest neighbour', kernel, caliper or local-linear matching (e.g., Diaz and Handa, 2006). The third step estimates the ATET as the mean difference in outcomes between treated and comparison observations. An approach with does not require matching with a separate group is the interrupted time-series design, although ITS is usually considered more credible where data are available on a contemporaneous comparison (controlled ITS) (Shadish et al., 2002).

There are, in fact, three main approaches to addressing confounding: prospectively through control groups (selected at pre-test using randomised allocation of treatment, where possible); retrospectively using statistical methods or direct control for observables using adjusted regression and matching or stratification (moderator analysis), which rely on the existence of untreated observations; or using mediator analysis, by collecting data on outcomes along the causal pathway which are orthogonal to confounders (Pearl and Mackenzie, 2018). The directive acyclic graph shows these approaches (Figure 4.7). There are many examples, cited above, of studies using control groups and statistical adjustment to address confounding. An example of an uncontrolled study using statistical adjustment and mediator analysis is Genser et al. (2008) (see also Barreto et al., 2007 and 2010) which presented outcomes along the causal pathway for a sanitation infrastructure

intervention to connect latrines to the public sewer. The outcome analysis included observed household hygiene behaviour (to account for transmission of pathogens in the private space), visible defaecation in the streets (to account for transmission in the public space), intestinal parasite measurement (helminth infections and giardiasis), and reported diarrhoeal disease.

A prominent example of an approach which manipulates a mediator variable, is the randomised encouragement design (e.g., Duflo et al., 2007). This approach used when an intervention is universally available, but information about it is not, uses instrumental variables to estimate the unbiased effect of starting and adhering to treatment (complier average causal effect), as used for example in studies of provision of credit for household water connections (Devoto et al., 2012; Ben Yishay et al., 2017). Although it is not usually presented as an example of orthogonal mediator analysis, it would appear to be since the encouragement is along the causal chain between intervention and outcome.

Figure 4.7 Three ways of addressing confounding



Source: author drawing on Pearl and Mackenzie (2018).

Covariate balance across treatment and control groups in RCTs is usually verified by presenting means and standard deviations of observable covariates, with or without tests for statistical significance (Bruhn and McKenzie, 2009). However, they should also present information about the randomisation process, specifically how random numbers were centrally generated (tossing a coin, drawing from a lottery, using a computer programme) and how allocation was concealed during recruitment of

participants (e.g., use of sealed, opaque envelopes in a medical trial), to ensure there was no subversion of the randomisation process (Higgins et al., 2011).

Non-randomised studies, on the other hand, need to argue convincingly, and present appropriate results of statistical verification tests, that the design or methods of analysis can account for unobservable and observable confounding. Data permitting, it is useful to make assessments of group equivalence at baseline according to observable covariates, along with statistical significance tests, under the assumption that these are correlated with unobservables. Factors which may invalidate group equivalence during the process of implementation, such as time-varying confounding, should also be considered in estimation, as they also should in RCTs of sustained interventions.

For example, the validity of PSM rests on two assumptions: overlap and unconfoundedness. There must be some degree of overlap of covariate distributions in treated and comparison to identify suitable comparisons, also called 'common support' (Gertler et al., 2010). Overlap is testable by assessing whether comparisons are identifiable for treated observations at extreme values of the propensity score function. For example, a cross-sectional evaluation of the impact of piped water connections to Indian households on reported diarrhoea, using National Family Health Service survey data, matched on individual characteristics (age, sex, education, religion and ethnicity of household head, assets, housing conditions) and village characteristics (infrastructure, educational and social infrastructure, state dummy variables) (Jalan and Ravallion, 2001). Following comparison of propensity scores, 650 unsupported treatment households were dropped from analysis from 8,827 treated households available out of a total sample size of 33,216 (Figure 4.8).

Unconfoundedness means comparison and treated units are on average identical to their matches on observable characteristics, with the exception that one group participates in the programme. In other words, unobservable characteristics affecting outcomes are assumed correlated with observables which are equally distributed across groups. In this respect, PSM has the same limitations as analysis of covariance (ANCOVA) or multivariate regression analysis, but it has two important advantages. Firstly, in PSM the

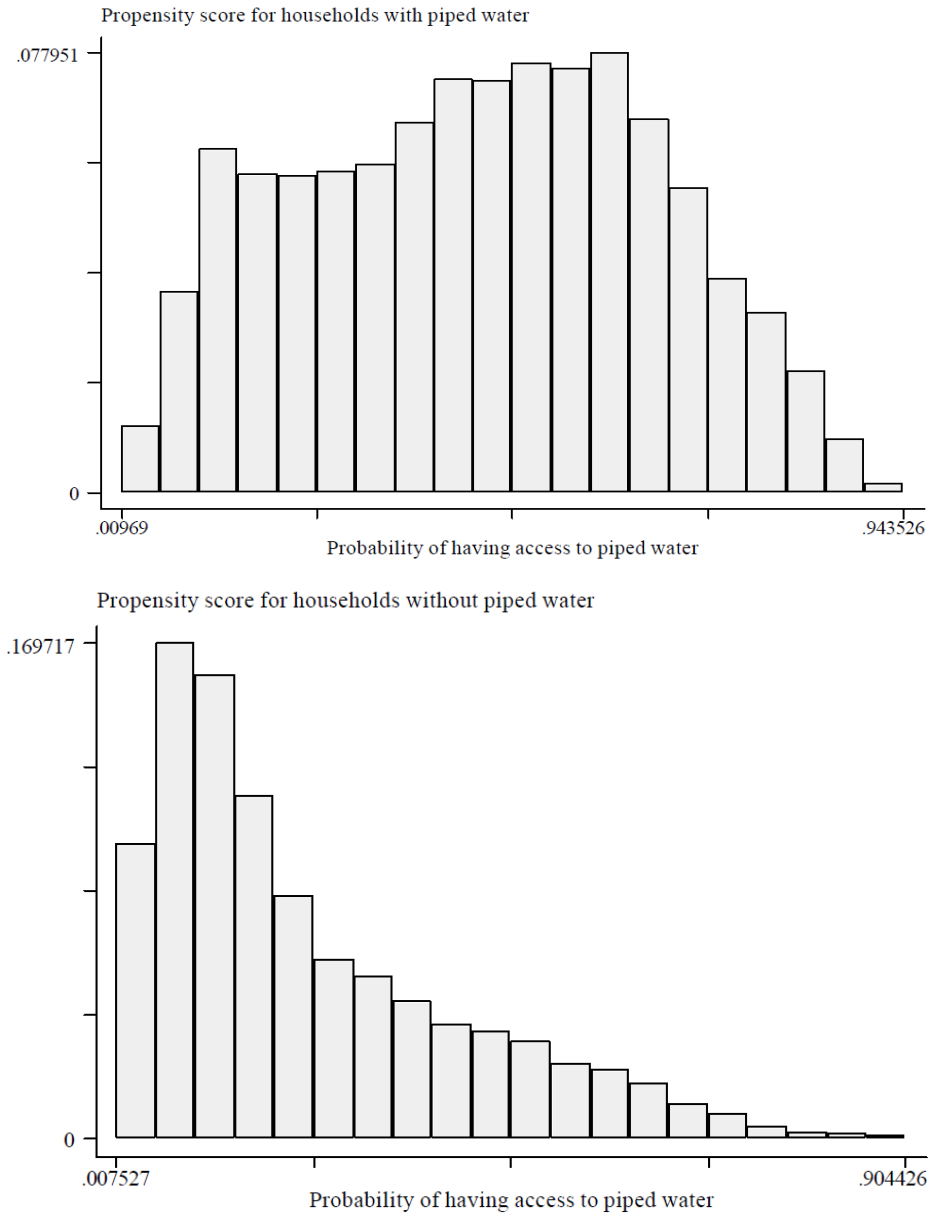
outcome is calculated non-parametrically, so it does not rely on any assumptions about the relationship between the outcome variable and covariates. Secondly, use of the matching algorithms improves comparability of observations (especially, in theory, those that weight paired observations according to their similarity like kernel matching), and therefore improves the accuracy of the estimated effect.[80] [81]

Unconfoundness is not testable and the existence or not of unobservable factors driving treatment needs to be assessed qualitatively, for example by checking that matching is done on as many pre-existing or time-invariant characteristics (including interactions between variables) as are available. Information on the process determining selection of treated units may also contribute to improving validity of the participation equation. Nevertheless, a falsification test for sensitivity of results to hidden bias exists (Rosenbaum and Rubin, 1982).

[80] An alternative technique is covariate matching (CVM) in which units are matched on individual covariates. CVM is likely to ensure that the treated and comparison units are more similar, as the match is not based solely on total probability scores but rather on each characteristic that affects participation. However, identification is more difficult as the more covariates that are considered, the greater likelihood of exclusion of observations in the treatment group, lacking common support, with a considerable loss of information. Coarsened exact matching (CEM), where matching is done on dichotomised continuous outcomes, aims to overcome this (Iacus et al., 2012), as used in evaluation of handwashing in India (Fan and Mahal, 2011).

[81] In large samples, the results do not vary with the choice of matching strategy. However, when the sample is small and overlapping limited, kernel and caliper techniques are preferred when some treated units have multiple close neighbours in the comparison group and others have only one. Additional concerns should be taken into account when using nearest-neighbour techniques in small samples. This guarantees a counterfactual for all the treatment units. However, in the presence of small samples, not all the counterfactual individuals might be identical to their matched individual in the treatment group. Therefore, when such techniques are used, it is necessary to assess whether the treated and comparison units are equal by comparing means or distribution of covariates. Caliper matching ensures overlap of matched observations but it can exclude observations in the treatment group when there are no comparison observations with a similar propensity score, resulting in a loss of information and potentially bias in ATET. Using replacement allows matching on more than one treated unit with the same comparison unit, with the effect of reducing sample size and inflating standard errors. Kernel matching reduces the loss of information by using the total sample size and weighting each match on the similarity of the propensity scores. Depending on the functional form assumed in the weighting, some observations might be excluded, although the loss of information is usually minimal.

Figure 4.8 Histograms of propensity scores for Indian households



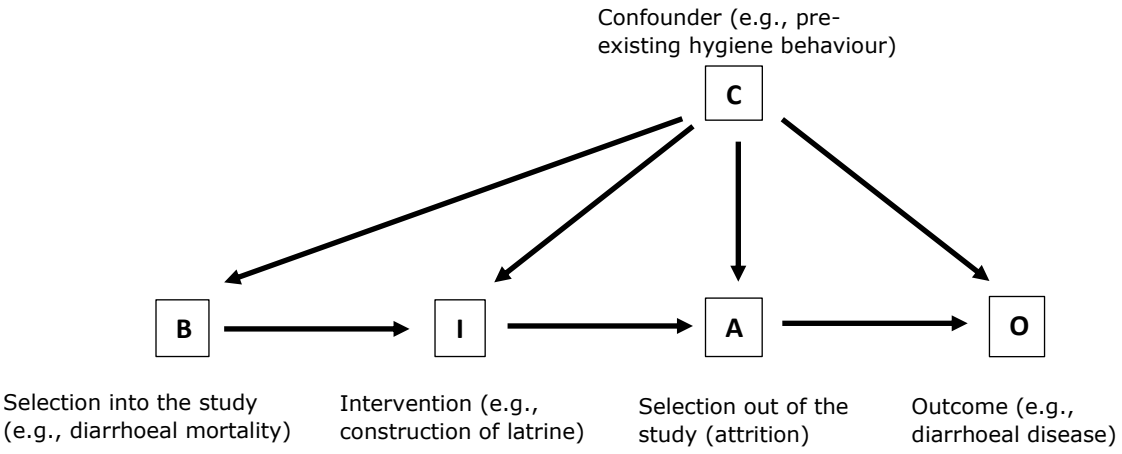
Source: Jalan and Ravallion (2001).

4.4.2 *Selection bias*

Selection bias occurs where some eligible treatment units or follow-up periods are excluded from data collection or analysis, which affects the observed relationship between intervention and outcomes. Using this definition, therefore, selection bias refers to selection into or out of study measurement – called 'sample selection bias' (Heckman, 1979). It can be differentiated from selection into treatment (self-selection or programme placement) which is defined under confounding above. The two main sources

are selection bias into the study – where, for some groups, all follow-up data are missing at pre-intervention stage or before outcomes start to be recorded – and selection bias out of the study – whereby post-intervention follow-ups are missing for some groups (Hernán et al., 2004). Selection bias is a special case of confounding and an important concern in retrospective studies and longitudinal studies like trials. Examples of selection bias, shown in the DAG (Figure 4.9), are non-random attrition in prospective studies (A) and censoring of eligible participants prior to treatment in retrospective studies (B) (e.g., diarrhoea or nutritional outcomes data are not available due to death of severely ill children).[82] The important point about selection bias is that it is problematic where it is differential between study arms (Briscoe et al., 1985; Sterne et al., 2016).[83] Where it is not differential, any threats to validity of the study would be threats to external validity.

Figure 4.9 Causal diagram showing selection bias



Source: author drawing on Swanson et al. (2017).

Biased selection into the study should not be problematic in prospective studies assigned at individual level, where full information about all eligible participants is available, and the process determining assignment is adequately concealed or non-manipulable by participants, practitioners or outcome assessors. It may be problematic in prospective studies assigned at group level, including cluster-RCTs, where those recruiting individual participants know about the treatment allocation status at group level

[82] Sterne et al. (2016) refer to this as inception/lead-time and immortal time biases.
[83] Briscoe et al. (1985) differentiated confounding, caused by variables distorting the relationship between the probability of exposure and the probability of illness, from selection bias, where variables distort the probability of exposure and the differential probability of reporting illness in treatment and control.

(Eldridge et al., 2008). It is also a potential source of bias in retrospective studies, including retrospectively designed RDDs, studies using Mendelian randomisation and other natural experiments (Swanson et al., 2017).

For example, analysis of the causal relationship between WASH practices and health indicators such as diarrhoeal illness and nutritional status should take account of selection bias due to mortality (Gómez et al., 1956) – that is, surviving children, on whom health indicator data are available, are not random draws from the underlying population, and survival may be determined by unobservable factors correlated with access to WASH and WASH practices (Lee et al., 1997). Where participants are recruited after treatment assignment in cluster-RCTs, or in retrospectively designed studies, an approach to resolve sample selection is Heckman's (1979) two-step procedure. The procedure adjusts the relationship between WASH treatment and health outcome by accounting for the missing observations in the lower part of the distribution. In the first stage the non-random selection variable, the probability of survival S_i, is estimated using probit estimation:

$$\text{Prob}(S_i = 1) = \Phi(\gamma \text{Wi}) \qquad (4.18)$$

where γ is the set of coefficients estimated on W explanatory variables and Φ indicates the cumulative normal density function. The inverse of Mill's ratio is calculated from the fitted values of the probit model, and included as an explanatory variable in a second-stage regression model of health status:

$$Z_i = \beta T_i + \lambda \frac{\phi(\gamma W_i)}{\Phi(\gamma W_i)} + u_i \qquad (4.19)$$

where Z_i is the health status of child i, ϕ is the probability density function of the normal distribution, λ the estimated coefficient on the inverse of Mill's ratio and u_i the error term incorporating unobservables not captured by the inverse of Mill's ratio.[84] It produces consistent estimates assuming that the error distribution of selection and regression equations is bivariate normal

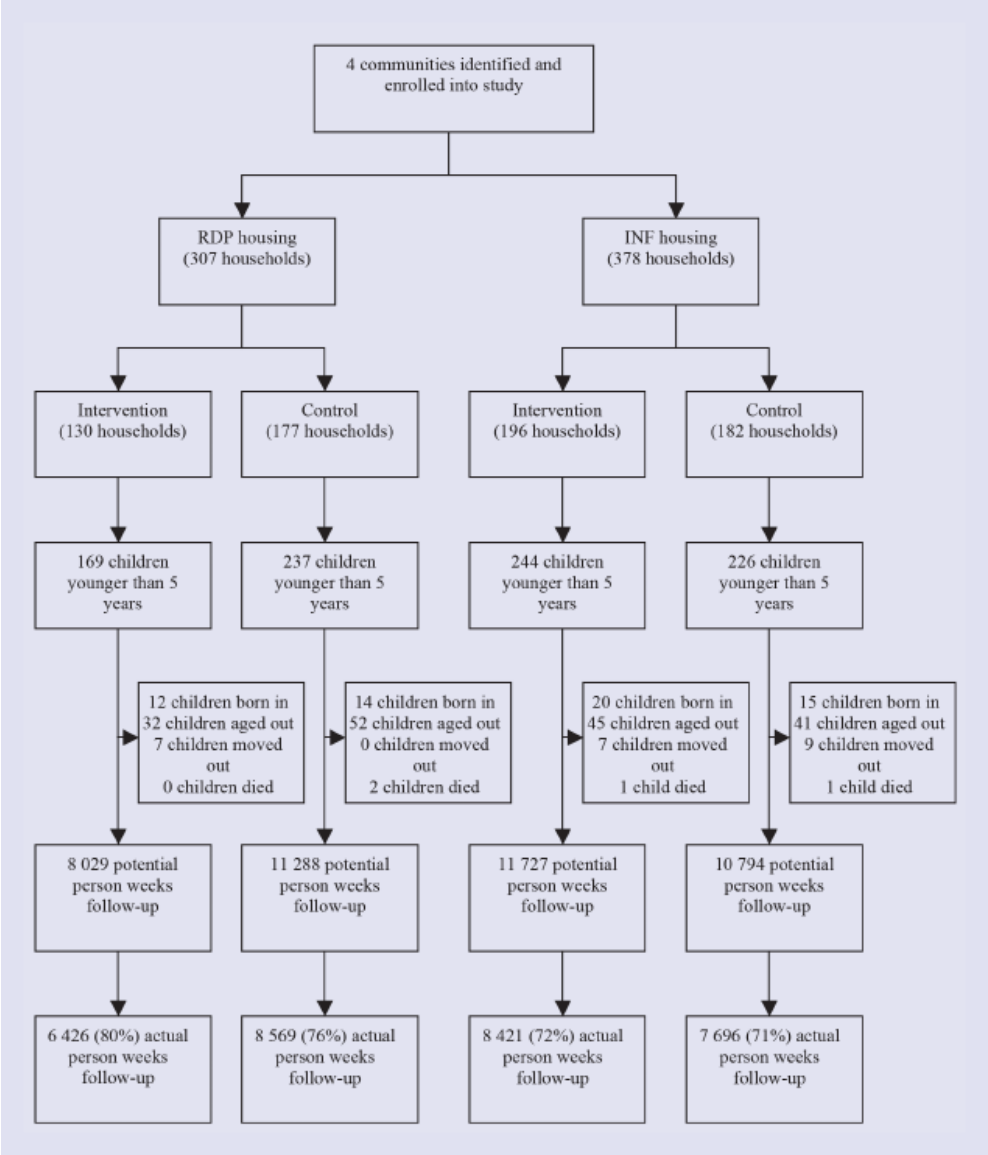
[84] The sign of the estimated coefficient on the inverse of Mill's ratio reflects the correlation between error terms in selection and regression equations, providing statistical evidence for non-random selection (Greene, 2002). In the case of morbidity and malnutrition, a negative coefficient is expected, indicating that the unobserved characteristics determining survival, such as preferences about childcare, broadly defined, are also negatively correlated with those influencing morbidity.

(Greene, 2002). Although the model is identified by the non-linearity of the first stage probit selection equation, the first-stage equation should also include an exogenous variable(s) determining survival but unrelated to child health status at treatment baseline (White et al., 2005). For example, the RCT by Luby et al. (2018) which found reductions in mortality in WASH treatment groups, but no effect on nutritional outcomes, could potentially use randomisation as an instrument in modelling the survival selection equation.

The method of assigning participants into treatment and control is an important concern in RCTs, where participants, practitioners or those recruiting participants can anticipate randomisation status before recruitment into the study and therefore affect who receives treatment. For example, where randomised allocation is not done centrally, or if done centrally, is communicated in open (unblinded) format (e.g., open random allocation schedule), it may be possible to subvert the randomisation process to affect the trials results. Similarly, in cluster assigned RCTs, bias may occur where recruiters of individual participants are not blinded to cluster assignment decisions and have incentives to choose participants based on that knowledge to affect the outcome of the trial (Eldridge et al., 2008). It is therefore important that randomised allocation is blinded until after recruitment of participants. Evidence for bias due to subversion of randomisation in meta-epidemiological studies, suggests odds ratios are of greater magnitude if randomisation is inadequately concealed (OR=0.83, 95%CI=0.74, 0.93; evidence from 102 meta-analyses, ratio of findings from 532 inadequately or unclearly concealed RCTs to 272 adequately concealed RCTs) (Wood et al., 2008).

Assessment is needed of the extent to which the design and methodology account for selection biases. The preferred approach is for authors to report the participant flow diagram. Figure 4.10 presents an example of a study of health impact evaluation of hygiene education in peri-urban South Africa (Cole et al., 2012). It reports numbers enrolled at baseline (pre-intervention), losses to follow-up, and reasons for these, by group. As is common in diarrhoeal disease research (e.g., Luby et al., 2004), the figure is, however, missing the reasons for person-weeks lost, which may be due to random missingness or due to factors relating to the outcome (e.g., severe gastro-intestinal infection requiring attendance at health facility).

Figure 4.10 Participant flow in a clustered non-randomised trial



Source: Cole et al. (2012).

Selection bias into the study in retrospective NRS may be addressed using selection models or inverse probability weighting (Hernán et al., 2004). Bias due to selection out of the study may be assessed by reporting losses to follow-up (attrition) by treatment group, the reasons for it by group, measuring the correlation with variables predicting outcomes, and the use of attrition-adjusted weights (Fitzgerald et al., 1998). Differential attrition (by treatment status) is considered more important than overall attrition, although analysis of both is needed. For example, Clasen et al. (2006, 2015) used overall attrition thresholds of 10 percent to differentiate low and high risk-of-bias studies. What Works Clearinghouse (Deke et al., 2015) attrition thresholds suggested that overall attrition can be as high as 50 percent as

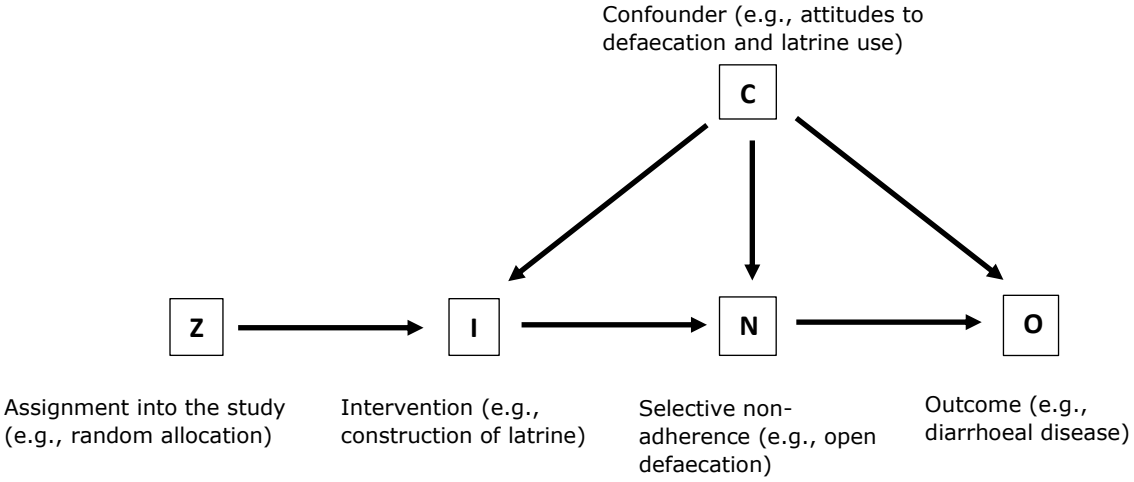
long as differential attrition was less than 1 percent, to produce consistent estimates. In contrast, the maximum acceptable rate of differential attrition was 6.3 percent provided overall attrition was 10 percent or less.

4.4.3 *Bias due to departures from intended interventions*

Participants receive a different intervention to the one intended due to 'performance bias' – no-shows, crossovers (contamination), spillovers, and implementation fidelity – and 'motivation bias' – e.g., Hawthorne and John Henry effects. No-shows and crossovers (also called switches) respectively occur where units assigned to treatment do not receive it and units assigned to control receive treatment. They are potentially problematic in all controlled studies including unblinded prospective trials, and in evaluations of sustained or adaptive intervention strategies (e.g., double blind RCTs with an adaptive design where participants cross over if they do not improve sufficiently). Assessment should be made of the extent to which potential bias due to switches is accounted for in design or analysis. It is usually handled using analysis of the effect of assignment to intervention using ITT analysis, or in analysis of the effect of starting and adhering to treatment (also called per-protocol analysis) using complier average causal effect (CACE) analysis.

Estimating the treatment effect in sustained interventions with non-compliance is problematic (Swanson et al., 2017). For example, in a trial of single-pit ventilated improved pit (VIP) latrine provision in rural areas, involving sustained behaviour change of participants, women in West Africa may self-select to not adhere, because of traditional rules around defaecating on top of their father-in-law's faeces; men may elect to defaecate in the open to avoid the cess pit filling up (in the absence of faecal sludge removal services); or children may be afraid to use it due to the dark (DAG in Figure 4.11) (Curtis et al., 1995). Information should therefore be collected on non-adherence to avoid bias in the ITT estimate (Clasen et al., 2014). However, attempts to estimate the per-protocol effect (starting and adhering to treatment) in sustained interventions with non-compliance (e.g., using CACE) is likely to be biased, which is particularly problematic in studies where ITT is not a viable estimand (e.g., retrospective studies) (Swanson et al., 2017).

Figure 4.11 Selection bias due to non-adherence of a sustained intervention



Spillovers, when members of the comparison group are exposed to treatment indirectly, through contact with treated individuals, are potentially problematic for all controlled studies (e.g., Miguel and Kremer, 2004). Cluster-level analysis may be required to address these sources of bias, and/or an assessment of the geographical or social separation of treated and untreated groups needed. For example, Ryder (1985) conducted a study in two islands without fresh water sources, one which received a new water supply system provided unlimited water supply, and the other a comparison in which women collected water daily from a stream on the largely uninhabited mainland, 1 mile away across open ocean. The islands were 6 miles apart. It seems very unlikely therefore that there would be any biases due to contamination of comparisons or spillover effects. In rural Bangladesh, Luby et al. (2018) used a buffer of 1 km around each enrolled cluster to reduce possibilities for spillover effects.

Another form of contamination, called 'substitution bias' (Heckman and Smith, 1995) may arise where controls obtain similar treatments from different providers (e.g., Maluccio and Flores, 2004). For example, it may be difficult to identify controls, particularly among 'donor darling' countries, due to existing coverage (co-interventions) by development partners (Figure 4.12). However, this problem may be overstated, since implementation is often variable, such as the neglect of hygiene promotion in WASH programming at scale (Jimenez et al., 2014). And even if relevant, it can be ameliorated, firstly through assignment at cluster level over a sufficiently large number to ensure balance in co-interventions on average across treated and untreated units, by collecting (and controlling for) information about co-

interventions, as done in the impact evaluation of *Red de Protección Social* (Maluccio and Flores, 2004) discussed in Chapter 5, Section 5.3. Others have called for estimation of dose-response relationships in evaluations where the unit of observation is the district, to address substitution effects (Victora et al., 2011).

Figure 4.12 Selected development partners working in Mozambique

Source: Victora et al. (2011).

Evidence from internal replication studies, discussed in Chapter 5, Section 5.3, indicates that matching is more effective in achieving balance when the observations being matched are geographically proximate (Diaz and Handa, 2006; Handa and Maluccio, 2010). At first glance, it might therefore appear that the possibilities of locating a good match and avoiding spillovers are mutually exclusive in programmes like hygiene campaigns or treatment of infectious diseases, where spillovers are expected. However, it is worth noting that this is less problematic when matching is done at group level, such as clusters of communities. Accurate matching at individual level that avoids bias in the effect estimate due to spillovers may necessitate matching individuals in different communities, or an assessment of the likelihood that the intervention or outcomes may spillover.

The other main source of deviation from intended intervention, motivation bias, is potentially more problematic in prospective studies, whether randomised trials or NRS. For example, the fact that monitoring participants influences their behaviours because they are aware of being watched, called Hawthorne effects in treated groups and John Henry effects (due to compensatory rivalry or resentful demoralisation) in controls (Bärnighausen et al., 2017b). 'Survey effects' or measurement as treatment, where being surveyed sensitises individuals to technologies like hygienic behaviour, promoting adherence among treated units or uptake among controls, has been observed in prospective impact evaluations using repeated measurement, including in WASH (Zwane et al., 2011). Bias due to survey effects may be less problematic in observational studies or in trials with fewer data collection follow-ups (Gaarder et al., 2011). It may therefore be useful to collect information on the frequency of measurement, to test for systematic differences in effects across studies (e.g., using moderator analysis), or if feasible to use additional study arms free from monitoring visits. For example, Banerjee et al. (2012) added a 'pure control' study arm, which was not informed about the project and only visited by investigators at the endline outcomes survey, to measure the motivation biases caused by monitoring of controls in a trial in Rajasthan, India. In Kenya, Null et al. (2018) had both an 'active control' (received visits by health promoters as well as data collection) and a 'passive control' (baseline and endline data collection only) to account for possible motivational biases.

There may be a trade-off between the need to monitor intervention fidelity and adherence of sustained interventions through frequent visits, and the need to reduce potential Hawthorne and survey effects. For example, it can be useful to evaluate intervention processes or adherence on a subsample of the treatment group (Fiala and Premand, 2017), or in selected treatment villages not enrolled in the impact evaluation, as done in an evaluation of the roll-out of the Indian government's *Swachh Bharat Abhiyan* total sanitation programme in Bihar (Dreibelbis et al., 2018).

4.4.4 *Bias in measurement of intervention and outcomes*

Bias in intervention measurement can be problematic where the explanatory variable is reported access to or use of WASH facilities. It is thought particularly unreliable where information about treatment status is obtained after implementation from participants who may have an incentive to misreport, or where recalling receipt of the intervention, or defining it adequately (e.g., its dose, frequency, intensity or timing), is difficult (Sterne et al., 2016). For example, as noted in Briscoe et al. (1985) people are more likely to report using better WASH facilities than they do. The effect of measurement error in the intervention, even under non-differential misclassification, is to bias downwards the treatment estimate (Newell, 1962; cited in Briscoe et al., 1985; Wooldridge, 2009).

Bias in measurement of outcomes due to recall and disease definition are potentially problematic in all studies where outcomes data are self-reported. For example, monthly recall was shown to lead to underreporting of acute illness and healthcare seeking behaviour in observational and experimental surveys in Delhi, compared to weekly recall (Das et al., 2009). The authors also found the bias was larger among poor respondents, for whom nearly 50 percent of illness and self-medication episodes and one-third of doctor visits were forgotten, which was consistent with illness being normalised in households whose sickness burden is higher.[85]

However, other sources of bias are potentially more problematic in prospective studies, whether randomised trials or NRS. These biases affect

[85] Briscoe et al. (1985) also discussed the effect of the normalisation of illness over time on the underreporting of diarrhoea among poor people.

measurement and therefore differ from performance bias which affects behaviour. Bias in outcomes measurement can be addressed in trials when participants or outcome assessors are blinded to intervention, or when outcomes are directly observed rather than self-reported. For example, social desirability (courtesy) bias may occur in open (unblinded) trials where participants are aware of treatment status and health outcomes data are measured by repeated self- or carer-reporting (Schmidt and Cairncross, 2009). This may occur for a number of reasons such as the desire to please the enumerator who is associated with the WASH intervention, or even due to survey participant fatigue due to repeated measurement, also called the 'bugger-off effect' (Clasen, 2013). Social desirability bias may also arise due to the 'Clever Hans effect', where participants are inadvertently induced to report favourably by outcome assessors (survey enumerators) (Beath et al., 2013). Even in the absence of an intervention (equivalent to double blinding), longitudinal measurement has been shown to cause participants to alter observed behaviours, where outcomes are easily modifiable at short notice (e.g., clean hands and presence of soap at handwashing station) (Arnold et al., 2015).

Double blinding participants and outcome assessors to intervention is usually impossible, with the exceptions of anti-bacterial hygiene interventions (Larson et al., 2004) and some household water treatment devices, although even this may be difficult due to water turbidity (Boisson et al., 2010). In some instances where double blinding is possible, it may not be approved in ethical review, for example, as reported in a study of chlorine disinfectant in Bangladesh (Ercumen et al., 2015a). Blinding of outcome assessors may be possible in trials of WASH technologies delivered to communities, as done by Pickering et al. (2015). In trials of household interventions, controls may be provided a 'placebo intervention' that does not affect outcomes of interest, e.g., children's books, notebooks, pens and pencils in a household water treatment and hygiene trial in Pakistan (Luby et al., 2004). In trials of community interventions, a number of possibilities for reducing misreporting of outcomes are possible. The following were implemented in a study of community-driven development in Afghanistan by Beath et al. (2013): outcome assessors were blinded to intervention status; respondents were kept unaware of the purpose of the survey in informed consent (see also Schmidt, 2014), and informed that responses would not determine the receipt of further assistance; intervention practitioners were

not informed about timing of survey or shown questionnaires in advance in order to avoid priming of participants; and questions were omitted from outcomes surveys that may have informed the treatment status or cued enumerator or respondent about the purpose of the survey. Blinding of clinical examiners to intervention status was done by Tadesse et al. (2017).

Another way to address this bias is through collection of 'placebo outcomes' which are in theory unrelated to the intervention. For example, Ercumen et al. (2015a) collected data on reported skin diseases and ear infections that would not be affected by the interventions (household water treatment through chlorination and safe storage) and Ercumen et al. (2015b) collected reported scrapes/bruises in the assessment of water supply reliability. However, the minimum condition for a placebo outcome is that it is not theoretically related to the intervention. For example, Ercumen et al. (2015b) and Augier et al. (2016) used respiratory illness in placebo analysis of water supply improvements. However, water supply availability may help reduce respiratory illness by enabling hand and domestic hygiene.[86]

Bias due to social desirability is less likely to be problematic in observational studies where participants do not associate data collection with a particular intervention. Evidence from meta-epidemiological studies suggests that unblinded studies of health interventions may be severely biased when outcomes are subjectively measured (e.g., reported by participants or practitioners); on average, pooled relative odds ratios (RORs) were of greater magnitude in unblinded studies (ROR=0.75, 95%CI=0.61, 0.82; evidence from 32 meta-analyses comprising 104 unblinded and 205 blinded trials) (Wood et al., 2008). However, when outcomes were observed, there were no differences between blinded and unblinded trials (ROR=1.01, 95%CI=0.92, 1.10; 44 meta-analyses of 210 unblinded and 227 blinded trials). Where outcomes measured all-cause mortality, there were no differences (ROR=1.04, 95%CI=0.95, 1.14; 18 meta-analyses of 79 unblinded and 121 blinded trials) (see also Savović et al., 2012).

As noted by Schmidt: "[t]he act of randomisation after informed consent when carried out at the household level almost precludes an unbiased

[86] Other methods proposed to elicit 'true' responses from participants include list experiments (Karlan and Zinman, 2012) and anchoring vignettes (King et al., 2004). These do not appear to have been used in WASH impact evaluations.

response in symptom-based questionnaire surveys" (2014, p.523). It follows that some of the issues around self-reporting in unblinded trials can be ameliorated in cluster-assigned evaluations of interventions provided at group level, where household consent is restricted to health outcome measurement not intervention delivery, so observation may appear unconnected to the intervention (Eldridge et al., 2008).[87]

4.4.5 *Bias due to selective methods of analysis and reporting*

Bias in reporting corresponds to selective reporting of outcomes (e.g., among multiple possible outcomes collected), selective reporting of results from sub-groups of participants, or selective reporting of methods of analysis (e.g., where multiple estimation strategies or specifications are used) (Rothstein et al., 2005; Sterne et al., 2016). There are usually thought to be two main sources of this bias. The first is significance inflation (or 'p-hacking') whereby researchers test multiple hypotheses until they find statistically significant results, which are then submitted for publication. The second source of selective reporting is therefore the non-significant findings from the published studies, as well as the non-significant findings from studies which ultimately remain unpublished, being left in the researchers' file-drawers (Rothstein et al., 2005; Ioannides et al., 2017; Vivalt, 2018).

These types of bias are particularly likely to be prevalent in retrospective evaluations based on observational datasets, where the method of analysis or outcomes are chosen based on results, but they are problematic in prospective studies as well. Presence of a study protocol (pre-analysis plan) can help determine the likelihood of bias, although it is recognised that many prospective (and nearly all retrospective) studies do not have or publish such plans,[88] nor is it possible to fully specify models for some methods like PSM

[87] For similar reasons, confounding bias due to the use of quasi-random processes to allocate groups are likely to be ameliorated in cluster designs, such as where schools or villages are alphabetised by name and then centrally assigned by alternation (Miguel and Kremer, 2004; Montgomery et al., 2016). In contrast, it is easier to see how individual alternation could be manipulated by participants or recruiters (e.g., those waiting in line, neighbouring households in a community).

[88] Trial registries have become common in recent years (e.g., clinicaltrials.gov; the American Economic Association RCT Registry http://www.socialscienceregistry.org; the Registry for International Development Impact Evaluations ridie.3ieimpact.org), aided by the refusal of some journals to publish impact evaluations without published trial registries or pre-analysis plans. In addition, some journals in economics (e.g., Journal of Development Economics)

and RDD in advance. Transparent reporting of any analyses that were determined post hoc may be undertaken. Reporting on all outcomes and participant sub-groups measured irrespective of findings may be helpful, although it is recognised that journals word limits may impede the author's ability to do this.

Vivalt (2018) examines effect sizes across RCTs and NRS on development programmes, suggesting that RCTs are less prone to 'significance inflation', as measured by bunching of p-values at the traditional significance level of 5 percent, than NRS. She also finds significance inflation for RCTs, particularly those done by economists and 'non-economists' (mainly health researchers) working in development research, to have diminished over time. In contrast, she finds biases from NRS to have increased over time. Some arguments cited for why this might be the case are: 1) greater competition among journals in some fields leading to only articles with significant findings being published; 2) the preference for RCTs in publication, over NRS, leading RCTs to be published regardless of findings; 3) the increasing requirement for registration of pre-analysis plans for RCTs; 4) the requirement for authors of RCTs to present unadjusted findings as standard, following the development of guidance like Consolidated Standards of Reporting Trials (CONSORT) standards for RCTs (Moher et al., 1998; Moher et al., 2010), reducing opportunities for these to be left in file-drawers; and 5) the 'equilibrium' in some fields or journals where p-hacking is more common, requiring researchers to engage in it to be competitive. It might also be thought that some journals or editors may have an incentive to report findings from more novel approaches yielding positive effects, where conduct standards have not been agreed in academia.

For example, in RDD it is standard for studies to report multiple specifications to check robustness. This includes testing the robustness of the results to the use of non-parametric methods using different bandwidths, use of weighting for matches further from the assignment threshold, and

have recently committed to publishing results of trials, whatever the findings, provided the study protocol is registered with them (Foster et al., 2018). The Journal of Development Effectiveness was the first development journal to encourage explicitly authors to submit null findings when it was established in 2009. As noted online: "The journal has an explicit policy of 'learning from our mistakes', discouraging publication bias in favour of positive results – papers reporting interventions with no, or a negative, impact are welcome". Available at: https://www.3ieimpact.org/resources/Journal%20of%20Development%20Effectiveness (accessed 23 March 2020).

functional form (e.g., step versus slope, linear or non-linear relationship between forcing variable and outcome) when modelling the relationship between assignment and outcome variables (Villar and Waddington, 2019)

Blinding to treatment status of researchers with potential reporting incentives, such as data analysts, is feasible, but rarely done. However, a study in Brazil blinded data analysts to intervention status in laboratory measurement of stool samples (Moraes et al., 2004), as also done by Emerson et al. (2004), Masset et al. (2011) and Stoller et al. (2011). Luby et al. (2018) also blinded data analysts, requiring two data analysts to conduct statistical analysis from raw datasets "with the true group assignment variable replaced with a re-randomised uninformative assignment" (p.e304).

In addition to assessment at study level, it is worth noting that selective methods of reporting can be tested for at the review level in analysis of small-study effects, which, under particular assumptions, are related to publication bias (Egger, 1997; Peters et al., 2007).[89]

4.4.6 *Adequacy of sample size*

A major issue in early WASH impact evaluations relates to the study sample being too small to estimate effects with statistical precision. Also referred to by Blum and Feachem (1983) as 'one-to-one comparison', this problem relates to the collection of data from dependent observations in a limited number of clusters, often from only one or perhaps two villages each in the treatment and comparator (e.g., Khan, 1987; Aziz et al., 1990). The issue is likely to be particularly problematic in the case of infectious diseases. Practically all statistical tests assume that each case is a statistically independent event. However, cases occurring in a single village or community cannot be considered independent (because people catch infections from one another).

Table 4.7 presents results from an evaluation conducted in two geographically separated informal settlements on the outskirts of Dhaka, Bangladesh (Khan, 1987). In Tongi, Oxfam had built five enclosed,

[89] Further discussion and analysis of publication bias in WASH studies is in Chapter 5 Sections 5.3 and 5.5.

communal pour-flush latrines with cemented seats, which drained into sedimentation tanks that stored sewage under anaerobic conditions to kill parasites. Most people reported using them, although young children defecated in their homes.[90] In Kalsi, the comparison area, each family, or group of families, shared an open, unlined pit latrine surrounded by a roofless bamboo enclosure, located next to their hut. Residents in both settlements were dewormed at the start of fieldwork to ensure comparability of health outcomes, and communities were found to be similar on observable characteristics (e.g., drinking water source, household size, literacy, occupation, and hut building material).

Table 4.7 Infectious disease in peri-urban areas of Dhaka: confidence intervals re-estimated for correlated observations

	All diarrhoea incidence	*Infant diarrhoea incidence*	*Hookworm prevalence*	*Giardia prevalence*
Tongi (communal latrines)	752/924	66/46	41/982	160/982
Kalsi (unimproved latrines)	579/823	66/44	19/807	171/807
Rate ratio	1.16	0.96	1.77	0.77
95% confidence interval	1.04, 1.29	0.68, 1.35	1.04, 3.03	0.63, 0.93
Design effect (*Deff*)	84	45	10	10
Adjusted 95% confidence interval*	0.01, 99.0	0.14, 6.74	0.11, 28.8	0.07, 8.86

Note: * adjusted for design effect assuming intra-cluster correlation coefficient equal to 0.1 for diarrhoea and 0.01 for hookworm and giardia. Data reported as number of cases per unit of population. Source: author using data presented in Khan (1987), Schmidt et al., (2010) and Schmidt et al. (2011).

The disease incidence findings are consistent with later policy guidance (e.g., WHO/UNICEF, 2000) about the limited protective effectiveness of communal and unimproved latrines. In this particular instance, it is likely that in Tongi, diarrhoea and hookworm would spread via water-washed transmission from person-to-person in the home, especially due to child

[90] Women may also prefer to not use communal latrines for safety reasons (Biran et al., 2011).

defaecation there, and by sharing communal sanitation facilities without adequate hand hygiene.[91] Households also bathed and washed in temporary ponds, which in Kalsi would likely be contaminated during the rainy season by the open, unlined pit latrines, which if accidentally swallowed would propagate giardiasis. However, the statistical findings, presented here assuming independence of observations, represent lower bounds of the correct standard errors and confidence intervals.

Where study participants are grouped into correlated clusters of observations, statistical calculations need to consider the design effect (e.g., Higgins and Green, 2011; Schmidt et al., 2011):

$$Deff = 1 + (m - 1)\rho \qquad (4.20)$$

where m is the average number of observations per cluster and ρ is the intra-cluster correlation coefficient (ICC). The corrected standard error calculation is simply the unadjusted standard error multiplied by the square route of the design effect (Waddington et al., 2012):

$$se' = se\sqrt{Deff} \qquad (4.21)$$

Although epidemiologists often assume *Deff* is 1.5 in diarrhoea studies (Victora et al., 1997), a review found that it was often much higher, ranging from 0.1 to 22 (Schmidt et al., 2011). Schmidt et al. (2011) presented design effect calculations for nine village- or neighbourhood-clustered studies, from which the most relevant ICC was $\rho = 0.094$, calculated from a study in urban Pakistan (Luby et al., 2005); this study was chosen as it most closely corresponded the example used here – an urban area with weekly visits over approximately one year to measure reported diarrhoeal disease.[92] Applying this to the data above yields a design effect of 84 for all age diarrhoea and 45

[91] In particular, hookworm spreads through contamination of the yard and communal defaecation areas, which improved sanitation, defined as the safe removal of faecal matter from the environment, provided to individual households, would be expected to decrease.

[92] The data are reported in Schmidt et al. (2011) who calculated *Deff* equal to 13.2 for community clustering in the Pakistan study. From this, together with the average number of observations per cluster reported by Schmidt et al. (2011) at 130, the author calculated the ICC at 0.094. While this may provide a useful approximation of ICC for diarrhoeal disease, it is likely to overestimate ICC for hookworm and giardia which were only collected twice (at start of fieldwork and endline). ICC=0.01 was therefore chosen for these outcomes drawing on examples in Schmidt et al. (2010), yielding *Deff* equal to 10 each for hookworm and giardia.

for infant diarrhoea, together with confidence intervals so wide as to suggest the findings are statistically meaningless.

4.5 External validity in impact evaluations

The value of an impact evaluation for policymaking depends on the rigour with which it is conducted (internal validity), and its relevance for application in different contexts and the units being investigated (external validity). Relatedly, construct validity is the external validity of the study to the intervention and outcome relationships it is attempting to measure. Some authors also use applicability and transferability to describe, respectively, the likelihoods that an intervention and study findings are relevant to a new, specific setting (Burchett et al., 2011). These all relate to the generalisability of the intervention and study findings (Table 4.8).

This section focuses on external validity. The external validity of a study depends on a range of factors, including the design, sampling frame, whether the intervention was implemented in 'real world' or controlled settings, use of theory, the context, intervention characteristics and duration of study (Bracht and Glass, 1968).[93]

Table 4.8 Concepts of relevance in impact evaluation

Concept	*Definition*
External validity	"[I]nferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes" (Shadish et al., p. 83). The extent to which the study has relevance to the 'real' world in which people are working (Bracht and Glass, 1968; Eisenstein et al., 2007).

[93] External validity may also refer to broader generalisability concepts (e.g., Green and Glasgow, 2006; Eisenstein et al., 2007). Analysis of effect estimates for sub-groups, such as gender, age, length of treatment, baseline prevalence and contextual factors can be used to explore external validity. In addition, drawing on a program theory can inform the understanding of heterogeneity by setting out hypotheses about the characteristics of contexts, populations and interventions likely to affect findings, that can then be tested empirically by additional data collection and sub-group analysis. So, too, can qualitative analysis and/or analysis of qualitative data where these are available. As noted by Campbell (1984, p.42) "[o]ur skills should be reserved for the evaluation of policies and programs that can be applied in more than one setting… The lack of this knowledge (whether it be called ethnography, programme history, or gossip) makes us incompetent estimators of programme impacts, turning out conclusions that are not only wrong, but are often wrong in socially destructive ways."

Construct validity	"[I]nferences from the sampling particulars of a study to the higher-order constructs they represent" (Shadish et al., p.65). This includes Type III errors (error in measurement of implementation or lack of implementation fidelity) and Type IV errors (e.g., outcome data collected are irrelevant for decision-making) (Scanlon et al., 1977; cited in Dobson and Cook, 1980).
Applicability	The likelihood that an intervention could be implemented in a new, specific setting.
Transferability	The likelihood that the study's findings could be replicated in a new, specific setting (i.e., that its effect would remain the same).

Source: Waddington et al. (2012).

At its narrowest conception, external validity refers to the treatment effect estimand produced by an impact evaluation, which is determined by the sample included in estimation, which itself relates to the relevance of the evaluation question. For example, RCTs estimate the average treatment effect (ATE) causal estimand for a population, whereas RDDs estimate the local average treatment effect (LATE) estimand for a (non-random) sample of treated observations.[94] Where there is non-compliance, the unbiased treatment effect estimator (ITT) gives the estimate of effectiveness of assignment to treatment. This is equal to the 'per-protocol' effect (the effect of starting and adhering to treatment) when treatment compliance (adherence) is perfect. When compliance is imperfect, ITT may be considered the relevant estimate from the perspective of a decision-maker considering implementation of a particular programme in the 'real world', where non-adherence is a factor determining implementation effectiveness (Bloom, 2006; White, 2014). 'Per-protocol' analysis gives the average treatment effect on the treated (ATET) estimand for adherents (whether in RCTs or NRS) and is therefore a measure of treatment efficacy (Eisenstein et al., 2007). Instrumental variables estimation can be used to estimate ATET for a (non-random) sample of treatment compliers, also known as the complier average causal effect (CACE) estimand.

[94] It is worth noting that the ATE estimated in an RCT is equal to the population ATE (PATE) if the sample recruited into the RCT is itself randomly selected. Usually, units are not recruited randomly into RCTs, hence RCT estimands are often also 'local' sample ATEs.

These quantities can be converted under the strong assumption of homogenous treatment effects across the sample (Table 4.9).[95] The ATET is equal to the ITT estimator – that is, the estimated difference in mean outcomes for treatment and control groups ($\bar{Y}_t - \bar{Y}_c$) – divided by the compliance rate for units allocated to treatment (T|Z=1). In other words, the treatment quantity is rescaled using only those who receive treatment as the denominator – that is, excluding no-shows.

Table 4.9 Treatment effect estimands under non-compliance

Estimand	*Effect size formula (mean difference)*	*Standard error of mean difference*
Intention-to-treat	$\bar{Y}_t - \bar{Y}_c$	$SE(\bar{Y}_t - \bar{Y}_c) = \sqrt{\frac{{s_t}^2}{n_t} + \frac{{s_c}^2}{n_c}}$
Average treatment effect on the treated	$\frac{\bar{Y}_t - \bar{Y}_c}{(\mathrm{T}\|\mathrm{Z}=1)}$	$\frac{SE(\bar{Y}_t - \bar{Y}_c)}{(\mathrm{T}\|\mathrm{Z}=1)}$
Complier average causal effect	$\frac{\bar{Y}_t - \bar{Y}_c}{(\mathrm{T}\|\mathrm{Z}=1) - (\mathrm{T}\|\mathrm{Z}=0)}$	$\frac{SE(\bar{Y}_t - \bar{Y}_c)}{(\mathrm{T}\|\mathrm{Z}=1) - (\mathrm{T}\|\mathrm{Z}=0)}$

Note: P proportion of total sample size $n = n_t + n_c$ allocated to treatment.
Source: author based on formulae in Bloom (2006).

For example, if the compliance rate (T|Z=1) in the treatment group were observed at 50 percent, for a homogenous effect (E) of a handwashing promotion intervention equal to a particular decrease in the rate of diarrhoea incidence – say, 30 percent, which is commonly found in meta-analyses of effects of handwashing – we would expect the ATET estimator to report a larger average effect. This would be calculated as: $E_{ATET} = E_{ITT}/(\mathrm{T}|\mathrm{Z}=1) = 30/0.5 = 60$ percent. The standard error of ATET is equal to the rescaled standard error of the ITT estimator $SE(\bar{Y}_t - \bar{Y}_c)$: the pooled standard deviation in outcome across treatment and control groups, SD(y), divided by (T|Z=1).

Instrumental variables estimation can be used to estimate ATET under the weaker assumption of monotonicity ('no defiers' due to no-shows or

[95] Evidence suggests this assumption may be unrealistic. For example, Oosterbeek et al. (2008) found different impact estimates for the two poorest quintiles in an evaluation of conditional cash transfers in Ecuador.

crossovers) to produce the complier average causal effect (CACE) estimand for a (non-random) sample of treated observations (Angrist et al., 1996; Angrist, 2004). CACE is also calculable using information on compliance (Bloom, 2006). It is equal to the ITT estimator divided by the difference in treatment receipt rate in treatment group (T|Z=1) and control group (T|Z=0), or the treatment quantity rescaled over those receiving treatment excluding no-shows and crossovers. For example, if the compliance rate (T|Z=1) in the treatment group is observed at 50 percent, and crossover rate (T|Z=0) observed at 10 percent, for an intervention effect of 5 percent, CACE is calculated as: $E_{CACE} = \frac{E_{ITT}}{[(\text{T}|\text{Z}=1)-(\text{T}|\text{Z}=0)]} = \frac{30}{0.5-0.1} = 75$ percent. Because of potential heterogeneity in treatment effect estimates over the population, ATET and CACE (and LATE) only generalise to treatment recipients and not necessarily to the sample or population average treatment effect (ATE).

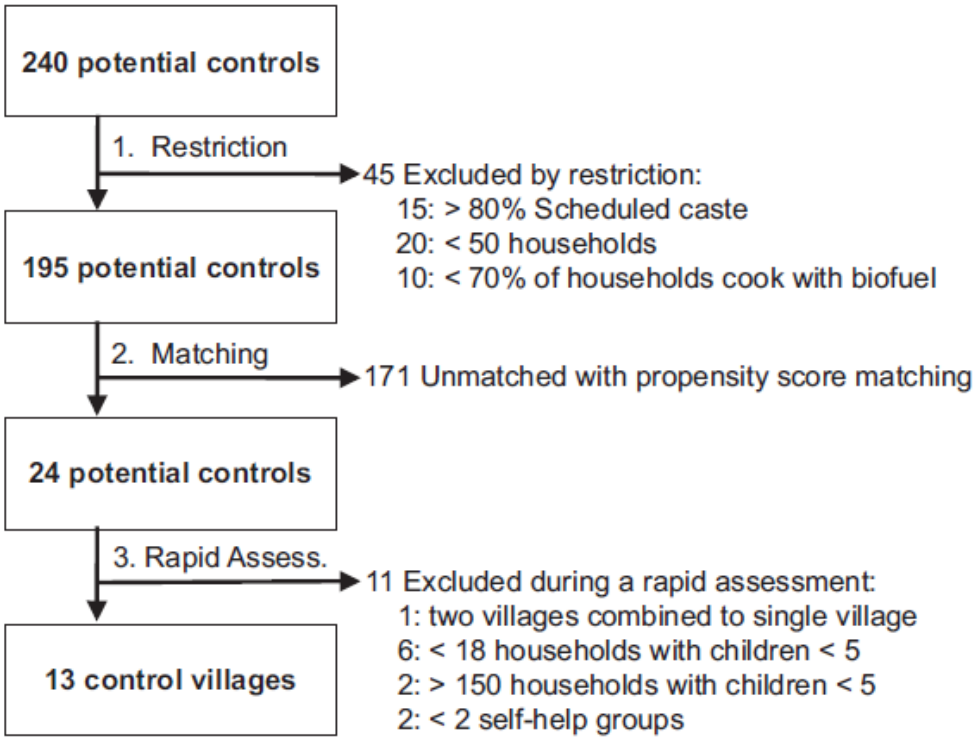
However, there are strong assumptions underpinning these calculations. Per-protocol analysis is only unbiased under the unverifiable assumption of unconfoundedness – that is, homogeneity of the treatment effect across all population units, including those lost to follow-up due to selection bias into or out of the study and self-selected non-compliers. CACE may be the treatment effect of interest for decision-making in NRS; for example, in an instrumental variables study of the effect of smoking on cancer using 'Mendelian randomisation' of genetic variants, the effect of interest will be the effect of smoking (estimated using CACE) rather than the effect of genetic inheritance (estimated using ITT). But CACE is difficult to define for sustained interventions due to issues of selection bias (Swanson et al., 2017), and in such cases, it is not clear whether the treatment variable in the instrumenting equation should be measured as a dichotomous variable or an ordered or continuous variable indicating degree of exposure (i.e., a dose-response).

The prediction interval discussed above (Section 4.2, Equation 4.13) is a useful concept, since it accounts for the greater uncertainty associated with unobserved variation due to these factors, when statistically pooling findings across studies. What may also be helpful in assessing external validity in single studies, although very uncommon in NRS, is in the reporting of study participant flow. For example, Arnold et al. (2010) reported participant recruitment in a cohort study of community water and hygiene programmes in Tamil Nadu, India, using statistical matching (Figure 4.13). Reporting of

participant flow is recommended in the CONSORT checklist for RCTs (Moher et al., 2010), and there is no good reason why they should not also be incorporated in prospective and retrospective NRS study reports to indicate reasons for dropouts (e.g., in statistical matching due to lack of common support).

The tension between internal and external validity is also used as an argument for conducting and using NRS in empirical policy research (Pritchett and Sandefur, 2013). Indeed, one of the reasons advocated for conducting NRS is that they do not disturb the usual processes of implementation, hence have greater relevance for decision-making (Bärnighausen et al., 2017a).

Figure 4.13 Selection process for propensity score matching



Source: Arnold et al. (2010).

However, even when compliance is perfect (or can be analysed adequately), there may still be issues in generalising the findings to the target population or intervention. For example, 'randomisation bias' may cause prospective evaluation programme participants to be systematically different from regular participants, for example where eligibility criteria are relaxed or participants are motivated change behaviour as a result of the threat of service denial (Heckman and Smith, 1995). The act of conducting a

prospective impact evaluation may lead to Hawthorne effects at the programme level due to the expectation of increased accountability, where policy makers make greater efforts to ensure the design of the intervention is suited for the implementation context, and/or practitioners are more careful in ensuring fidelity of implementation. A related issue is what may be termed 'evaluation placement bias': impact evaluations are more likely to be undertaken of programmes that are more effective, in circumstances more amenable to their successful implementation (Pritchett and Sandefur, 2013). Schmidt notes ruefully: "[i]t is difficult to escape the conclusion that the literature on the impact of water, sanitation and hygiene is unreliable in its entirety, and in any case, it only represents results from those trials and studies that are feasible – they would not be there otherwise" (2014, p.524).

4.6 An approach to assess bias comprehensively in randomised and non-randomised studies

Risk-of-bias tools were reviewed according to the extent to which they identified evaluation criteria and signalling questions for non-randomised approaches used in impact evaluations (Waddington et al., 2017). That review included tools aiming to assess RCTs and NRS together (Downs and Black, 1998; Cochrane Effective Practice and Organisation of Care (EPOC), undated;[96] Hombrados and Waddington, 2012; National Institute for Health and Clinical Excellence (NICE), 2009; Reisch, 1989; Sherman et al., 1998; Scottish Intercollegiate Guidelines Network (SIGN), 2011; West et al., 2002). We also included tools aiming to appraise only non-randomised studies (Cowley, 1995; Effective Public Health Practice Project (EPHPP), undated; Kim et al., 2013; Sterne et al., 2016; Wells, undated). Selected tools providing comprehensive internal and external validity assessments (Valentine and Cooper, 2008) and those focusing on external validity (Green and Glasgow, 2006; Montgomery et al., 2013) were also assessed.[97]

[96] The EPOC tool was developed drawing on the Cochrane risk of bias tool (Higgins et al., 2011).

[97] Green and Glasgow (2006) presented a tool to evaluate the potential for generalisation of effectiveness research, defined as "attempts to study programs under typical, rather than optimal conditions" (Green and Glasgow, 2006, p.127). They grouped categories under reach (e.g., is the intervention participation rate among the target group reported?), representativeness (e.g., is comparison made of the similarity between study setting and programme setting?), implementation fidelity (e.g., are data presented in level and quality of implementation of different components?), programme mechanisms (are data reported on processes or mediating variables through which programme achieves effects?), outcomes (are

The assessment indicated that existing tools contained evaluation criteria for domains of bias that are relevant to RCTs and NRS with selection on unobservables. However, most of the tools were not designed to assess causal validity of these studies, meaning that the 'signalling questions' on which biases are evaluated were not sufficiently targeted, particularly in the domains of confounding and reporting biases. For example, randomisation (sequence generation and allocation concealment) was usually the only method proposed to account for unobservable confounding. No single tool fully evaluated the internal validity of quasi-experimental designs and natural experiments. Of particular concern was the lack of a comprehensive risk-of-bias tool for *a priori* credible designs, in particular natural experiments. For example, four tools presented signalling questions for RDDs (Valentine and Cooper, 2008; Schochet et al., 2010; Hombrados and Waddington, 2012; Chief Evaluation Office, undated), of which Hombrados and Waddington (2012) included questions on all relevant domains of bias addressed here.

Most tools that aimed to assess randomised and non-randomised studies did not enable consistent classification of both approaches, or of different NRS methods, across the same evaluation criteria (e.g., NICE, 2009). Sterne et al. (2016) ask assessors to consider an unbiased 'target trial' (Hernán et al., 2016) against which a given NRS should be compared. This approach has arguably been useful in getting reviewers from outside of the clinical trials community to think about sources of bias which they may previously have been unaware. However, there are also instances where trials may be biased in ways which are not applicable to observational studies (e.g., performance bias due to Hawthorne and John Henry effects, as noted above). Application of these instruments may therefore lead to inappropriate risk-of-bias assessment for NRS, especially natural experiments with selection on unobservables.

---

these relevant for guidelines or policy, including quality of life, and are potential negative consequences and moderator analyses for sub-groups of participants reported?), and maintenance (e.g., are data reported on sustainability of programme implementation and effects at least 12 months following treatment, and analysis made of representativeness of attritors?). Eisenstein et al. (2007) discussed a comprehensive approach to measuring implementation fidelity, drawing on design (by planners), delivery (by implementers), uptake (by participants) and contextual factors that may moderate these aspects (e.g., socioeconomic status). This was later developed into the Oxford Implementation Index (Montgomery et al., 2013).

The tool presented here in Appendix A, built on the bias domains and signalling questions in existing tools, in particular those articulated by Sterne et al. (2016) and a critical appraisal tool that was previously developed by the author and a colleague (Hombrados and Waddington, 2012), and the review of WASH impact evaluations contained in this chapter. Based on the review of existing critical appraisal tools, signalling questions were developed for the four main areas of bias: confounding and selection bias; bias due to departures from intended interventions (performance bias and motivation bias); bias in measurement of intervention and outcomes; and bias in selection of the reported result. This is presented as an integrated assessment tool, covering randomised and non-randomised studies, incorporating both the study design and its execution in analysis. Risk-of-bias assessment is based on what is reported regarding the assumptions of the designs and the methods with which they are addressed (Littell et al., 2008).

The tool follows the principles for risk-of-bias tools by Higgins et al. (2011) – in particular, bias domains and signalling questions being chosen using both theoretical and empirical considerations. For example, signalling questions drew on findings from internal replication studies (see Chapter 5) about those characteristics under which NRS are able to produce comparable estimates to RCTs, specifically when information about the programme allocation approach was known, when baseline characteristics were incorporated (including baseline measures of the outcome), when matched cases were geographically local (Cook et al., 2008) and where 'rich controls' were used (e.g., Handa and Maluccio, 2010).

Higgins et al. (2011) also called for risk-of-bias tools not to use quality scales. Evidence suggests it is not appropriate to determine overall bias using weighted quality scales (Jüni et al., 1999; Herbison et al., 2006). Authors of critical appraisal tools have instead shown that it is possible to assess overall bias based on transparent decision criteria. Finally, according to Higgins et al. (2011) judgment in assessments is required to reach decisions. While this may be necessary in some instances, specific reporting requirements are indicated (e.g., around the use of confirmation and falsification tests) to ensure as much consistency across users as possible. For completeness, the approach also incorporates statistical precision and external validity, although questions about the latter are primarily sought for subsequent

analysis, rather than being incorporated into critical appraisal (i.e., risk-of-bias) judgements.

Recognising the importance of having information about the programme assignment in adequately modelling selection, the first section of the tool asks the user to clarify what information is known about the treatment allocation mechanism at group and individual levels in the study – in particular, whether the approach to treatment allocation is rationed by supply (e.g., individual or group targeting) or demand driven (participant self-selection) (Appendix A Table A1). Clarity is also sought on whether the independent variable in the study measures provision of an intervention, or an exposure (e.g., access to a particular WASH technology). If an intervention study, it is necessary to assess whether it is a baseline intervention (e.g., administration of deworming tablet) or continuous intervention (e.g., provision of hardware or software technology requiring behavioural modification). Questions relating to implementation processes are also raised at the outset, including information about implementation fidelity, programme take-up and adherence among participants.

The assessor is then asked to clarify whether the intervention allocation was controlled by researchers (e.g., through randomisation, discontinuity assignment, statistical matching), policymakers or practitioners (e.g., lottery, individual or household means-testing, community or geographic targeting), or participants (self-selection). Information is sought on the methods used to address confounding (e.g., randomisation, DD, ITS, RDD or other statistical method) and the sample used in estimation of the treatment effect (e.g., whether this represents the ATE or LATE).

Following Section 4.5 above, information is also sought about external validity (Appendix A Table A1):

- Study length (follow-up period) and number of follow-ups.
- Sampling frame for the study, and sampling approach at cluster and individual levels (whether random or purposive).
- Inclusion of a programme theory, and collection of data on outputs, intermediate and endpoint outcomes (causal pathway analysis).
- Intervention design and implementation (whether by researchers, policymakers or practitioners).

- Intervention scale: whether the study is a trial, pilot study or small-scale project (e.g., implemented in a few villages by researchers), or a programme evaluation (e.g., implemented at province or national scale by government, private sector or an NGO).

Part 2 of the tool relates to the specific bias domains (Appendix A Table A2):

1) *Confounding (bias domain 1)*: baseline characteristics are similar in magnitude, unbalanced characteristics are controlled in adjusted analysis; for randomised approaches, adjustments to the randomisation were considered in the analysis (e.g., stratum fixed effects, pairwise matching variables); time-varying confounding such as differential adherence in sustained interventions.
2) *Selection bias into the study (bias domain 2)*: randomisation approach and allocation concealment for individual and cluster-randomisation. For non-randomised studies, timing of follow-up.
3) *Attrition (selection bias out of study) (bias domain 3):* total attrition and differential attrition across study groups (presentation of average characteristics across treatments and comparisons, and reasons for losses to follow-up). In cluster designed studies, where respondents are not followed over time, assessment is needed of the sampling strategy.
4) *Departures from intended interventions due to motivation bias (bias domain 4)*: observational data versus experimental data with clear link to intervention (informed consent); repeated measurement (frequency and regularity of survey rounds); Hawthorne, John Henry effects, and survey effects.
5) *Departures from intended interventions due to performance bias (bias domain 5)*: no-shows and crossovers, addressed using ITT or CACE; spillover effects addressed through geographical distance between treatment and comparison; differential contamination by external programs (treatment confounding) addressed through information about adherence behaviour.
6) *Measurement error (bias domain 6)*: length of recall, definition of intervention and outcome, timing of data collection (seasonality, or seasonal variation accounted for some other way), method of data collection (observed versus reported), blinding of outcome assessors and, where possible, participants.
7) *Analysis reporting bias (bias domain 7)*: pre-analysis plan or study protocol, reporting outcomes as indicated in methods, reporting ITT alongside other estimators (if relevant), blinding of data analysts.

8) *Unit of analysis error (bias domain 8)*: methods used to adjust standard errors to account for correlation of observations within clusters (e.g., cluster-robust standard errors).

Some of the signalling questions used to operationalise evaluation of bias are design-specific, most obviously for confounding and reporting domains (for which Appendix A Table A2 bias domains 1 and 7 distinguish RCTs, RDDs, DID, and IV). However, RCTs, NRS, prospective and retrospective studies may have different *a priori* risks of selection bias, performance bias and measurement error, which are incorporated into the tool. For example, prospective studies (randomised and non-randomised trials) require assessment of Hawthorne, John Henry and survey effects under motivation bias. Cluster-RCTs and retrospective NRS (natural experiments and purely observational studies) require assessment of selection bias into the study. Retrospective NRS need careful assessment of measurement of the intervention.

Risk of bias due to confounding in RDDs includes questions about the definition of the assignment scale (continuous or discrete), the specification of the relationship between assignment and outcome, treatment confounding, and the assessment of balance. Thus we might expect credible RDDs to: use a continuous variable for assignment; use an appropriate method to examine the relationship with outcomes (e.g., non-parametric kernel regressions) as well as report sensitivity analysis; report a graph of the discontinuity to show no other discontinuities in the assignment variable within the window of interest; report a histogram (kernel density plot) of the assignment variable to spot bunching around the threshold which might be indicative of manipulation; and report baseline data to assess the pre-intervention relationship. For reporting bias, papers would be expected to present multiple findings for all outcomes using multiple bandwidths, preferably pre-specified.[98] The final section of the tool asks the user to clarify the units of analysis, treatment and (if relevant) randomisation (Appendix A Table A2 bias domain 8).

A limitation of risk-of-bias approaches is that they may unintentionally foster suppression of information, over reporting information non-favourable to

[98] Signalling questions for RDDs were developed in a separate review (Villar and Waddington, 2019).

the study. To address this limitation, reviewers are encouraged to down-grade studies that do not report information necessary to validate a particular bias domain (e.g., participant flow, method of randomisation, placebo tests, and so on). Following Sterne et al. (2016), each signalling question may score 'yes', 'probably yes', 'probably no', 'no' and 'unclear'. 'Unclear' is listed after 'no' to indicate that it is the lowest score attainable, so that studies are not penalised for reporting more comprehensive information, even if that undermines the assumptions of the approach. An explicit decision rule then links responses to signalling questions to a decision about risk of bias: 'low risk of bias', 'some concerns', and 'high risk of bias'. For example, total attrition is nearly always reported, differential attrition by study group less so, and only the most comprehensive studies report reasons for attrition by study group, or group-wise correlation between attrition and sample characteristics. Hence, the RCT of a conditional cash transfer programme reported in Maluccio and Flores (2005), which reported all of these characteristics about attrition, was awarded 'some concerns' in Chapter 5 Section 5.3. The same study, previously reported in Maluccio and Flores (2004), omitted to report differential attrition and the correlation with sample characteristics. It would therefore have also been awarded 'some concerns' due to missing information.

It is possible that "strong researcher involvement in implementation" (as used in risk-of-bias assessment by Brody et al., 2016, p.36) could also be considered as a threat to internal validity, since it might increase the likelihood of Hawthorne or John Henry effects, survey effects or courtesy bias in reported outcomes. However, the likelihood of these biases is assessed through number of survey rounds, types of data collected in outcomes survey, and whether outcomes are observed or reported. In these circumstances, therefore, a well-conducted trial with strong researcher involvement in implementation may still result in an unbiased intervention effect, but the external validity of the results may be questionable, since it may have little relevance to intervention delivery in the 'real world' (Bracht and Glass, 1968).

Because of the judgement required, risk-of-bias assessment is usually done by multiple coders independently (for at least a sample of the primary studies reviewed). Inter-rater reliability for risk-of-bias assessments were undertaken in two systematic reviews undertaken by the author on farmer field school (FFS) (Waddington et al., 2014) and participation, inclusion,

transparency and accountability (PITA) (Waddington et al., 2019). Two statistics are commonly used to assess the reliability of judgements made by different raters: the percentage agreement and kappa. The simple percentage agreement is the number of cases which received the same rating, p_0 , divided by the total number of cases rated, N. Cohen's (1960) kappa κ adjusts the simple percentage agreement to take into account the share of agreed ratings that would be expected by chance alone p_e , calculated from the number of individual cases n that are rated k by raters 1 and 2, n_{k1} and n_{k2} respectively.[99] The formulae for these measures together with their standard errors are in Table 4.10. They assume each rater's coding was done independently of the other's.

Table 4.10 Inter-rater agreement

Estimator	*Formula*	*Standard error*
Percentage agreement	$\frac{p_0}{N}$	$\sqrt{p_0\ (1-p_0\ )}$
Cohen's kappa	$\kappa = \frac{p_0\ - p_e}{1-p_e}$ where $p_e\ = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$	$\sqrt{\frac{p_0\ (1-p_0\ )}{\mathrm{N}(1-p_e\ )^2}}$

Percentage agreement, expected agreement and kappa were calculated for both reviews (Table 4.11). In Waddington et al. (2014), there were minor disagreements in ratings for confounding and departures from intended interventions. The main disagreement was for blinding, which was downgraded from 'low risk of bias' after peer review, as the "Campbell Collaboration peer reviewer disagreed with the positive assessment [that had been given] due to lack of blinding of outcome assessors and data analysts" (p.48), hence relevant studies that had not used blinding were reallocated to 'high risk of bias'. In all cases, however, the results were broadly in agreement and all agreements were statistically significantly different from expected agreements.

[99] In addition, Cronbach's alpha α can be applied when judgements are made on a scale, such as the implicit Likert-scale used in generating an overall risk of bias assessment from the individual bias domains: $\alpha = \frac{k}{k-1}\left(1 - \sum_i \frac{s_i^2}{s_t^2}\right)$, where k is the number of items contributing to an overall score, s_i the standard deviation across raters of the scores for each item and s_t the standard deviation across raters of the overall scores (Cronbach, 2004). Alpha was considered to lack statistical value in this case because the risk of bias tool provides clear guidance on how to calculate an overall rating.

Waddington et al. (2019) assessed inter-rater reliability for a sample of 14 studies included in a review of PITA. The review included a broad range of studies and, unlike Waddington et al. (2014), incorporated both RCTs and NRS. The areas of bias where inter-rater agreements were not adequately reached were departures from intended interventions and motivation bias. It is worth noting that motivation bias was measured under 'deviations from intended intervention' in Waddington et al. (2014), which was limited to non-randomised studies mainly done retrospectively, as no RCTs of the FFS approach had been undertaken at that point. In such cases, departures from intended interventions are largely a result of spillover effects due geographical proximity of intervention and comparison groups, since motivation bias is not considered problematic in retrospective studies.[100]

In contrast, Waddington et al. (2019) used an almost identical approach to that presented in this Thesis, where departures from intended interventions may arise due to non-compliance and motivation bias due to Hawthorne effects, as well as spillover effects. The low kappa scores and lack of statistically significant differences between percent and expected agreement, suggest that it was difficult to assess these bias domains consistently. More objective questions that are less subject to judgement are needed for these two domains, which will also depend on the topics being reviewed (e.g., whether the intervention or outcome measured is communicable matters for spillovers; motivation bias is less problematic for objective outcomes).[101]

A final word is warranted on the utility of a combined risk-of-bias score across all categories. This relates to the relationship between bias in a particular domain on the estimated treatment effect. For example, lack of control for confounding would usually be expected to increase the effect size. Similarly, unconcealed allocation or attrition, causing selection bias, and reported outcomes may also increase the effect size. Motivational effects due to repeated measurement may either increase effects (Hawthorne effects, John Henry effects due to resentful demoralisation, survey effects), or reduce them (John Henry effects due to compensatory rivalry, 'bugger-off effects').

[100] The review drew on the first draft of the tool presented here, where selection bias and attrition were subsumed into the confounding domain (Hombrados and Waddington, 2012).
[101] These factors were incorporated into the risk of bias assessment used in Chapter 6 of this Thesis.

In contrast, deviations from intended interventions due to spillovers (contamination) or no-shows and crossovers (switches) are likely to reduce the estimated effect size. These points are further discussed in Chapter 5, Section 5.2.

Table 4.11 Inter-rater assessment in two reviews that used the tool

	Waddington et al. (2014): FFS			*Waddington et al. (2019): PITA*		
	Percent agreement	*Expected agreement*	*Kappa (P>|z|)*	*Percent agreement*	*Expected agreement*	*Kappa (P>|z|)*
Confounding	98%	52%	0.95 (0.000)	64%	29%	0.53 (0.001)
Selection bias	-	-	-	64%	23%	0.50 (0.000)
Attrition bias	-	-	-	93%	26%	0.90 (0.000)
Performance bias	90%	38%	0.85 (0.000)	21%	36%	-0.22 (0.919)
Motivation bias	-	-	-	50%	50%	0.00 0.500)
Outcome measurement	100%	62%	1.00 (0.000)	57%	28%	0.41 (0.000)
Analysis reporting	100%	38%	1.00 (0.000)	71%	34%	0.57 (0.000)
Blinding	21%	8%	0.14 (0.000)	71%	27%	0.61 (0.000)
Observations	42			14		

Note: - indicates score not available in Waddington et al. (2014), where selection bias and attrition bias were measured under 'confounding' and motivation bias was assessed under 'deviations from intended intervention'.

Source: author.

Chapter 5 Systematic evidence on bias from study replication in international development

5.1 Introduction

This chapter explores the relationship between probable bias, on the one hand, drawing on implementation of the critical appraisal approach presented in Chapter 4, and empirical estimates of bias on the other. Theory is ambiguous as to whether randomised and non-randomised studies typically produce reliable treatment effect estimates, or whether probable bias, determined by risk-of-bias assessment, is correlated with the deviation in findings from the 'true' value. Furthermore, the assumptions underpinning non-randomised study designs, as well as those underpinning the implementation of RCTs (e.g., selection bias and attrition), are untestable. Their verification therefore rests on empirical replication.

Results from two empirical analyses of bias are presented, in order to address Thesis Question 3 on the extent that biases predicted in theory are reflected in empirical estimates. The chapter draws on existing approaches to compare a given estimator (whether from a randomised or non-randomised study) with an unbiased, causal benchmark estimator, which is usually considered to be the estimate produced by a well-conducted RCT (Bloom et al., 2002). Section 5.2 presents a review of international development systematic reviews incorporating RCTs and NRS, which critically appraised risk of bias using the approach outlined in Chapter 4. This approach uses 'cross-study' comparison (or external replication) of effect sizes from randomised and non-randomised studies, selected using systematic search methods and pooled using meta-analysis.

However, cross-study comparisons are primarily indirect comparisons from studies conducted among different underlying populations. They may therefore be subject to confounding due to context, population, intervention, and so on. In contrast, internal replication studies use 'within study' comparison of a particular estimator with a causal benchmark, both of which

are sampled from the same target population. A systematic review of international development impact evaluations using non-randomised internal study replication of randomised trials was therefore undertaken to quantify bias, from which heuristics on bias in different study designs and methods of analysis were developed and incorporated into the critical appraisal tool. The results are reported in Section 5.3.

5.2 Review of international development systematic reviews

Cross-study comparisons, also called meta-epidemiological studies (Sterne et al., 2002), are used to examine whether study findings from external replication vary systematically according to methodological characteristics; for example, whether randomised trials are more or less likely to report bigger effects than non-randomised studies (e.g., Sacks et al., 1982; Concato et al., 2000; Kunz and Oxman, 1998). A cross-study comparison of interventions in social psychology, containing a very broad range of study designs (Lipsey and Wilson, 1993), found the point estimates calculated from meta-analyses of NRS with contemporaneous comparison groups were similar on average to those from RCTs. In contrast, studies using uncontrolled pre-test post-test designs were likely to produce estimates almost two-thirds bigger than controlled designs.[102] In a recent review of over 15,000 effect size estimates from 635 papers on international development programme effects, Vivalt (2020) also found that "RCTs do not exhibit significantly different results than quasi-experimental studies within an intervention-outcome combination" (p.32).

5.2.1 *The relationship between study methods and the magnitude of effect*

One might expect NRS to lead to bigger effects than RCTs for two main reasons: 1) non-adherence in trials; and 2) publication bias. Non-adherence in trials causes the intention-to-treat estimator, which is the unbiased estimate from the RCT, to be smaller than treatment-on-the-treated, which is the estimate usually produced by NRS data analysis. Publication bias

[102] Using the distance metric notation presented below in Section 5.3.4, the absolute standardised mean difference between NRS and RCTs was $\left|\hat{d}_{NRS} - \hat{d}_{RCT}\right| = 0.46 - 0.41 = 0.05$ standard deviations. The absolute standardised mean difference between controlled comparisons and uncontrolled comparisons was $\left|\hat{d}_{UC} - \hat{d}_{CC}\right| = 0.76 - 0.47 = 0.29$ standard deviations.

causes NRS estimates to be typically larger than RCTs because the research and publication process enables RCTs to be published more easily, regardless of the study's findings, all else equal.

However, it is not clear *a priori* whether studies with lower probability of bias – measured as across confounding, selection bias, performance bias, measurement error or reporting bias domains – are likely to estimate effects that are systematically different from studies with higher risk of bias (Lipsey and Wilson, 1993; Kunz and Oxman, 1998). For example, the mean difference between findings from a survey of meta-analyses of studies with high and low methodological quality ratings was estimated to be only 0.03 standard deviations (Lipsey and Wilson, 1993).

To take specific examples of bias domains, lack of control for confounding is sometimes expected to increase the effect size, but adding control variables may also increase it. This may be demonstrated using the standard formula for estimating omitted variable bias (Wooldridge, 2009):

$$\tilde{b}_T = \beta_T + b_x\tilde{\delta}_1 \qquad (5.1)$$

where $\tilde{b}_T$ is the coefficient for treatment variable T estimated using the (mis-specified) model omitting covariate X, b_x is the coefficient estimate for covariate X from the (correctly specified) model including both variables, and $\tilde{\delta}_1$ is the covariance between T and X. Whether $\tilde{b}_T$ is larger or smaller than the true coefficient estimate from the correctly specified model, β_T, depends on the product of the signs of the relationship between omitted variable and outcome β_x and the correlation between the treatment and covariate, determining $\tilde{\delta}_1$ (Table 5.1).

Table 5.1 Differences in estimated coefficients due to confounding

	$Corr(T,X) > 0$	$Corr(T,X) < 0$
$\beta_x > 0$	$bias = \tilde{b}_T - \beta_T > 0$	$bias = \tilde{b}_T - \beta_T < 0$
$\beta_x < 0$	$bias = \tilde{b}_T - \beta_T < 0$	$bias = \tilde{b}_T - \beta_T > 0$

Source: Wooldridge (2009, p.91).

For example, a cross-national observational regression study estimated the effect of latrine provision on diarrhoea but was not able to incorporate

several potential confounders such as socioeconomic status and water access (Esrey, 1996). Excluding measurement of socioeconomic status would cause overestimation of the effect of latrines, since socioeconomic status would be expected to be positively correlated with both diarrhoeal disease and latrine access and use; the same argument also holds for omission of water supply use and functioning from the model (Cairncross and Kolsky, 1997). To take another example, randomised trials frequently include covariates in outcomes estimation, whether to account for imbalance in baseline characteristics, or to improve precision of estimation by reducing the unexplained component in the regression equation (mean squared error) (Bloom, 2006). The anticipated effect of inclusion of covariates, such as carers' education and observed hygiene practices, in a trial of water and sanitation with non-adherence, would be to inflate the estimated effect towards the treatment on the treated estimate. This is because these covariates would be expected to be positively correlated with the diarrhoeal disease outcome and the omitted variable measuring adherence.

This concept is closely related to selection bias, which is a special case of confounding (Heckman, 1979; Chapter 3, Section 3.4.2 of this Thesis). Where selection bias of participants into the study is positively correlated with the outcome but negatively correlated with treatment, then the estimated treatment effect would be underestimated. For example, in a context of high rates of child diarrhoeal mortality, a cross-sectional study of access to safe child excreta disposal (potties) might underestimate the effect on diarrhoea morbidity and nutrition (Lee et al., 1997). This is because the sample of children measured in the study is not randomly selected but censored by mortality. Hence, in an observational study, the observed distribution of treatment outcomes includes those children who benefit from the potties through lower morbidity, those who benefit through survival, but who may have higher rates of morbidity, whereas the observed distribution of control outcomes excludes those who died who are also likely to have worse sanitation access.[103] Selection bias may therefore lead to estimation of perverse effects (i.e., a negative effect of safe disposal on morbidity) because of differential selection into treatment and comparison group study arms. Similarly, if selection bias out of the study (attrition or losses to follow-up) were positively correlated with the outcome and higher in the treatment

[103] Sterne et al. (2016) refer to this as inception/lead-time and immortal time biases. Wooldridge (2009, p.323) calls this 'endogenous sample selection'.

group (i.e., positively correlated with treatment), the effect size estimate would increase; whereas if attrition were higher in the control group (negatively correlated with treatment) then the effect for attrition positively correlated with outcome would be to decrease the estimated effect.

Performance bias, or departures from intended interventions due to spillovers (contamination) would tend to reduce the mean difference between treatment and control outcomes, for an effective intervention, while deviations due to no-shows and crossovers (switches) would also reduce the estimated effect measured using ITT.

It is not clear *a priori* whether the act of participating in a prospective study (whether randomised or not) is likely to lead to systematic differences in effects from retrospective studies. For example, motivational bias due to repeated measurement may either increase effects (Hawthorne effects, John Henry effects due to resentful demoralisation, survey effects), or reduce them (John Henry effects due to compensatory rivalry, 'bugger-off' effects). Others have argued that intervention fidelity may be better in trials (what might be called the Hawthorne effect of monitoring the trial on intervention practitioners) than routine practice, leading to bigger effects on average (Kunz and Oxman, 1998). Whether or not this is true, due to their costs and profile, it is highly likely that trial sites are chosen in favoured circumstances with the best chances for desirable outcomes. in contrast, where self-selection bias or programme placement bias are likely to lead to those in routine practice (i.e., non-randomised allocation) with better prognostic factors receiving treatment, and those with worse prognostic factors being allocated to comparison, randomised studies would be expected to have smaller effects on average.

Measurement error in the outcome variable is usually expected to cause bias in estimation when it is systematically related to one of the explanatory variables, such as the treatment (Wooldridge, 2009). For example, where outcomes are self-reported we would only expect systematic bias in the effect estimate if the data were collected differently in treatment and control group (e.g., different numbers of follow-ups), or there were different incentives affecting accuracy of reporting in each group (e.g., treatment units over-report to gain repeated treatment and control units underreport to gain treatment). Where incentives are not differential, there is lesser expectation

of systematic bias, even for outcomes collected using unreliable methods such as self-report (provided the recall period was sufficiently short for accuracy). Therefore, it might reasonably be thought that measurement error in outcomes is more likely to be problematic in prospective studies, where differential incentives are more likely to operate due to incentives, than in retrospective studies. However, Briscoe et al. (1985) also showed that even under non-differential misclassification, effect estimates are biased towards zero where outcomes are measured with error, and the bias increases the more frequent its incidence. Bias may therefore be expected to be lesser for less common outcomes, for example death as found in meta-epidemiological analyses (Wood et al., 2008; Savović et al., 2012). Wood et al. (2008) also found that studies using reported outcomes estimated bigger effects than measured outcomes in unblinded trials, with the exception of all-cause mortality. Measurement error in the explanatory variable is classically thought more problematic than measurement error in the outcome (Wooldridge, 2009). This is because, whether differential or not, it causes bias towards zero in the parameter estimate. It is commonly thought problematic in WASH field research. For example, Briscoe et al. (1985) noted that "those who use poor facilities will tend to report using better facilities" (p.13).

Bias in reporting is problematic in all studies, but likely to be especially problematic in NRS due to the publication process (Vivalt, 2020). For example, it is likely to be easier to publish an RCT without specification searching (p-hacking) to obtain a significant effect, particularly in journals that require trial pre-registration as a condition of publication. In contrast, few NRS are published that do not find significant effects; registration and pre-analysis planning, or encouragement of null findings, is almost unheard of for retrospective studies.

In summary, it is not clear *a priori* that bias, for individual domains or overall, would exert a systematic effect on either inflating or deflating treatment estimates, although we would expect a systematic effect on inflation of the variance of the effects both within and between studies. Perhaps the only prediction possible is that, in a meta-analysis containing biased estimates, and therefore deviation from the true effect in either direction – whether systematically in one direction or not – we would expect greater variance in the pooled effect.

5.2.2 *Analysis of systematic reviews that used the risk-of-bias tool*

The author reviewed international development systematic reviews that have used the tool outlined in Chapter 3. The reviews were selected purposively as those Campbell reviews which had used the tool.[104] Table 5.2 lists the reviews, the types of studies included, and the bias domains used in critical appraisal. The reviews covered a broad range of topic areas including agriculture, education, economic development, governance and women's empowerment. Systematic review authors were therefore encouraged to modify the tool to incorporate domains of bias that they considered most relevant for the literature. Some domains of bias were considered more widely applicable than others. Thus, all reviews assessed bias due to confounding and selectively reported analysis. Most reviews assessed selection bias, but only via selection into intervention (e.g., self-selected participation) or selection out of the study (attrition bias), and not usually selection of participants into the study itself (selection bias as defined in this Thesis). Nearly all reviews measured departures from intended intervention, even if that was in some cases restricted to spillover effects. Not all reviews assessed motivation bias (e.g., due to Hawthorne and survey effects), particularly those where prospective designs were not included. All reviews assessed outcome reporting bias, but this was restricted in several cases to selective reporting of outcomes (i.e., file-drawer effects) rather than bias in the methods used to collect outcomes data (e.g., whether outcomes were reported or observed) (Oya et al., 2016; Piza et al., 2016; Molina et al., 2016; Ton et al., 2017; Stone et al., 2019). A number of reviews also included 'other risks of bias' such as the use of recalled baseline data (e.g., Vaessen et al., 2014), similarity of data collection over time (Carr-Hill et al., 2016), missing data other than attrition (e.g., imputation) (Baird et al., 2013), coherence of results (e.g., Brody et al., 2015), and strong researcher involvement in implementation of the intervention (Chinen et al., 2017; Stone et al., 2019).

[104] These are all systematic reviews published by the Campbell Collaboration International Development Coordinating Group (IDCG); the author was the senior editor for the reviews included here. Some of the systematic reviews used an earlier version of the tool discussed in this Thesis (Hombrados and Waddington, 2012), others combined the tool with other approaches (e.g., Sterne et al., 2014, which was the precursor of Sterne et al., 2016).

Table 5.2 Systematic reviews and meta-analyses using critical appraisal tool

Authors	*Sector*	*Outcomes*	*# RCTs included*	*# NRS included*	*Domains of bias assessed*	*Other biases assessed*
Baird et al. (2013)	Education	School attendance	15	27	Confounding, attrition bias, departures from intended intervention, outcome measurement, reporting bias	Missing data, recalled baseline
Brody et al. (2015)	Micro-finance	Women's empowerment	5	18	Confounding, attrition bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Coherence of results, recalled baseline
Carr-Hill et al. (2016)	Education	School drop-out, test scores	9	17	Confounding, attrition bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Missing data, similarity in data collection over time
Chinen et al. (2017)	Vocational training	Women's employment, earnings	26	9	Confounding, attrition bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Researcher involvement in intervention implementation
Hemming et al. (2018)	Agriculture	Adoption of practices, agricultural yield, farm income	2	13	Confounding, attrition bias, departures from intended intervention, intervention and outcomes measurement, reporting bias	
Lawry et al. (2014)	Agriculture	Agricultural income	0	20	Confounding, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Missing data
Molina et al. (2016)	Governance	Health outcomes	10	5	Confounding, departures from intended intervention, outcomes measurement, reporting bias	Recalled baseline, blinding
Oya et al. (2017)	Agriculture	Income, wages, schooling	0	43	Confounding, attrition bias, departures from intended intervention, motivation bias, outcomes measurement, reporting bias	
Piza et al. (2016)	Vocational training and finance	Firm performance, employment creation	6	23	Confounding, departures from intended intervention, outcomes measurement, reporting bias	Recalled baseline, blinding
Samii et al. (2014a)	Agriculture	Environment conservation, poverty	0	11	Confounding, departures from intended intervention, motivation bias, intervention and outcome measurement, reporting bias	

Authors	*Sector*	*Outcomes*	*# RCTs included*	*# NRS included*	*Domains of bias assessed*	*Other biases assessed*
Samii et al. (2014b)	Agriculture	Environment conservation, poverty	0	8	Confounding, departures from intended intervention, motivation bias, intervention and outcome measurement, reporting bias	
Stone et al. (2019)	Education	Literacy	9	7	Confounding, attrition bias, departures from intended intervention, motivation bias, outcomes measurement, reporting bias	Small sample size, researcher involvement in intervention implementation
Ton et al. (2017)	Agriculture	Agricultural yield	0	22	Confounding, attrition bias, departures from intended intervention, motivation bias, outcomes measurement, reporting bias	Coherence of results, recalled baseline, similarity in data collection over time
Tripney et al. (2013)	Vocational training	Employment, hours worked, income	3	23	Confounding, attrition bias, departures from intended intervention, outcome measurement, reporting bias	
Vaessen et al. (2014)	Micro-finance	Women's economic empowerment	4	21	Confounding, attrition bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Coherence of results, recalled baseline, similarity in data collection over time
Waddington et al. (2014)	Agriculture	Adoption of practices, agricultural yield, income	0	93	Confounding, attrition bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Coherence of results, blinding
Waddington et al. (2019)	Governance	User engagement, provider response, access to services, use of services, attitudes to services, wellbeing, relationship with state	19	16	Confounding, selection bias, departures from intended intervention, motivation bias, outcome measurement, reporting bias	Blinding

Source: author.

Findings from meta-analyses of these studies were extracted to conduct analysis of study design and risk-of-bias categories. In several cases, insufficient details of meta-analysis were reported in the papers, most commonly because reviews did not report moderator analysis by study design and risk of bias (Baird et al., 2014; Carr-Hill et al., 2016; Chinen et al., 2018; Oya et al., 2017; Piza et al., 2016; Waddington et al., 2019). For example, the meta-analyses in Piza et al. (2016) were not reported separately for RCTs and NRS. In contrast, Baird et al. (2014), Carr-Hill et al. (2016), Chinen et al. (2018) and Waddington et al. (2019) reported meta-analyses according to study design but did not further disaggregate risk-of-bias status for RCTs and NRS separately.[105] In these four cases, it was necessary to extract the study-level data reported in the papers, or obtain the datasets from the authors, and re-analyse the findings. In two instances there were insufficient numbers of included studies to conduct analysis of effect sizes (Molina et al., 2016; Stone et al., 2019).

There are several issues in using meta-analysis to compare the implementation of the critical appraisal tool across reviews. The first is that the effect sizes may be computed differently. Most problematically, there are differences in the effect size used across meta-analyses. For example, some reviews in education reported standardised mean differences (SMDs) (e.g., Petrosino et al., 2012) while others reported odds ratios (ORs) (e.g., Baird et al., 2014). Some reviews in agriculture used response ratios, which are applied to continuous variables to measure the treatment mean as a proportion of the control mean (Waddington et al., 2014). It is possible to convert between SMD and OR (Sańchez-Meca et al., 2003), or to estimate OR from the risk ratio, by assuming a mean risk in the control group (Higgins and Green, 2011). However, there is no natural way to compute between response ratio and SMD, other than by recalculating effect sizes. In this review of pooled effects, all OR were transformed into SMD using methods given in Appendix C. Less problematically, effect sizes may be calculated to measure positive and negative outcomes differently. For example, a reduction in diarrhoeal disease (a positive outcome) may be measured as an odds ratio less than one, whereas other positive outcomes, such as increases

[105] Waddington et al. (2019) synthesised a large number of outcome variables. Prior to meta-analysis, the author compiled these into broad outcome constructs (citizen engagement, provider engagement, access to services, use of services, attitudes to services, wellbeing and attitudes to the State) using the method of synthetic effects to ensure that each meta-analysis only included independent effect sizes (given in Appendix C equation A38).

in good health practices, are measured as odds ratios greater than one. Where necessary, therefore, effect sizes were inverted so that positive outcomes were measured as odds ratios greater than one (SMD>0), and negative outcomes as odds ratios less than one (SMD<0).

A final potential problem is that there are multiple formulae to calculate effect sizes (Appendix C) which may not yield the same values of SMD. For example, the effect size calculated from the test statistic of an adjusted treatment effect regression is likely to be greater in magnitude than the equivalent effect size calculated from the same data using group means and pooled standard deviation. However, this was deemed less problematic since effect sizes within reviews are more likely to use consistent methods, and the purpose of the review was primarily to compare within-review pooled effect sizes (that is, pair-wise comparisons of pooled effect sizes composed of different study designs or risks of bias, presented in the same review).

Stata software was used to estimate meta-analytic pooled effects in this and subsequent chapters (Palmer and Sterne, 2016). Figure 5.1 overlays the distribution of effect sizes for RCTs and NRS contained in the meta-analyses. There are two main points of interest. Firstly, the modes of both distributions exceeded zero, suggesting most development interventions have positive effects. Secondly, the peaks of the distributions of effect sizes are the same for both types of design, which may be suggestive of equivalence in effects. However, the distribution of NRS pooled effects is skewed further to the right than that for RCTs, indicating that variance of pooled effects for NRS is bigger, as expected. A similar analysis by risk of bias suggests similar modes but greater variation for pooled effects comprised of studies with higher risk of bias (Figure 5.2). However, the pooled effect size data on which the comparisons are based come from different meta-analyses with different sample sizes and within-meta-analysis variances. It is therefore worth exploring more formally whether effect sizes are systematically different for RCTs and NRS.

In order to facilitate comparison of pooled effects by design and risk of bias, the distance estimate was calculated for each meta-analysis (Lipsey and Wilson, 1993). Distance, defined as equal to the difference in pooled standardised effects $\widehat{D}_{NRS}$ for i pair-wise pooled effect comparisons of NRS

and RCTs corresponding to the same intervention-outcome meta-analysis, was calculated as:

$$\widehat{D}_{NRS_i} = \hat{d}_{NRS\ i} - \hat{d}_{RCT\ i} \qquad (5.2)$$

where $\hat{d}_{NRS}$ is the estimated pooled effect size for non-randomised studies and $\hat{d}_{RCT}$ is the estimated pooled effect size for RCTs. A difference greater than one indicates that the pooled effect for non-randomised studies is larger than that for RCTs. The distance was calculated for 39 NRS effect sizes subject to *e* risk of bias, where *e* may equal 'low risk', 'medium risk' or 'high risk':

$$\widehat{D}_{NRS}{}^{e}{}_{i} = \hat{d}_{NRS}{}^{e}{}_{i} - \hat{d}_{RCT\,i} \qquad (5.3)$$

Similarly, the distance in pooled standardised effects between 'low risk of bias' RCTs and other RCTs was calculated for 22 effect sizes reported in the reviews:

$$\widehat{D}_{RCT}{}^{e}{}_{i} = \hat{d}_{RCT}{}^{e}{}_{i} - \hat{d}_{RCT}{}^{u}{}_{i} \qquad (5.4)$$

where $\hat{d}_{RCT}{}^{u}$ is the estimated pooled standardised effect for 'low risk of bias' RCTs and $\hat{d}_{RCT}{}^{e}$ the estimated pooled effect for medium and 'high risk of bias' RCTs. The standard errors of the differences were calculated assuming independence:[106]

$$se(\widehat{D})_i = \sqrt{s^2_{NRS_i} + s^2_{RCT\,i}} \qquad (5.5)$$

where *s* is the standard error of the pooled effect size for each study design comparison.

[106] Independence is a reasonable assumption given that each pair-wise distance estimate was calculated from two pooled effects drawing on different studies.

Figure 5.1 Number of pooled effects by study design

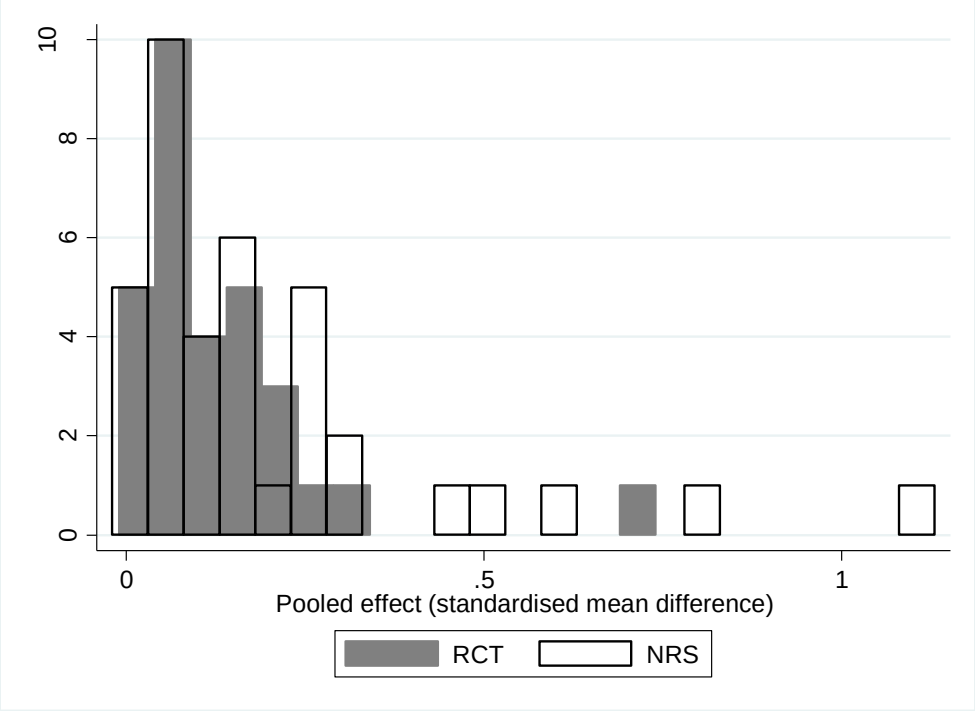
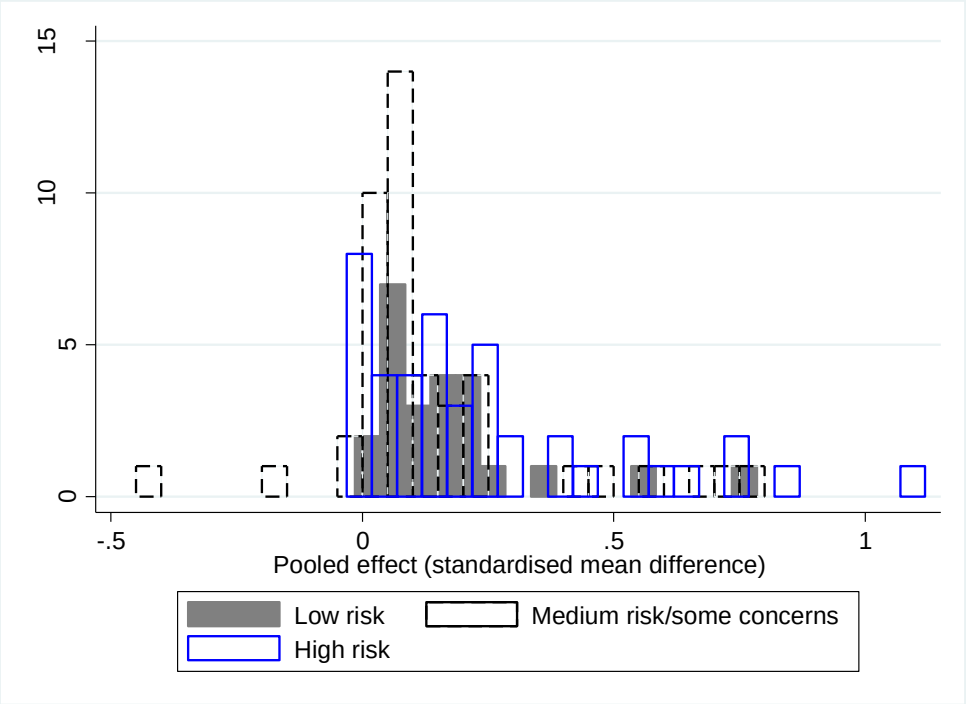
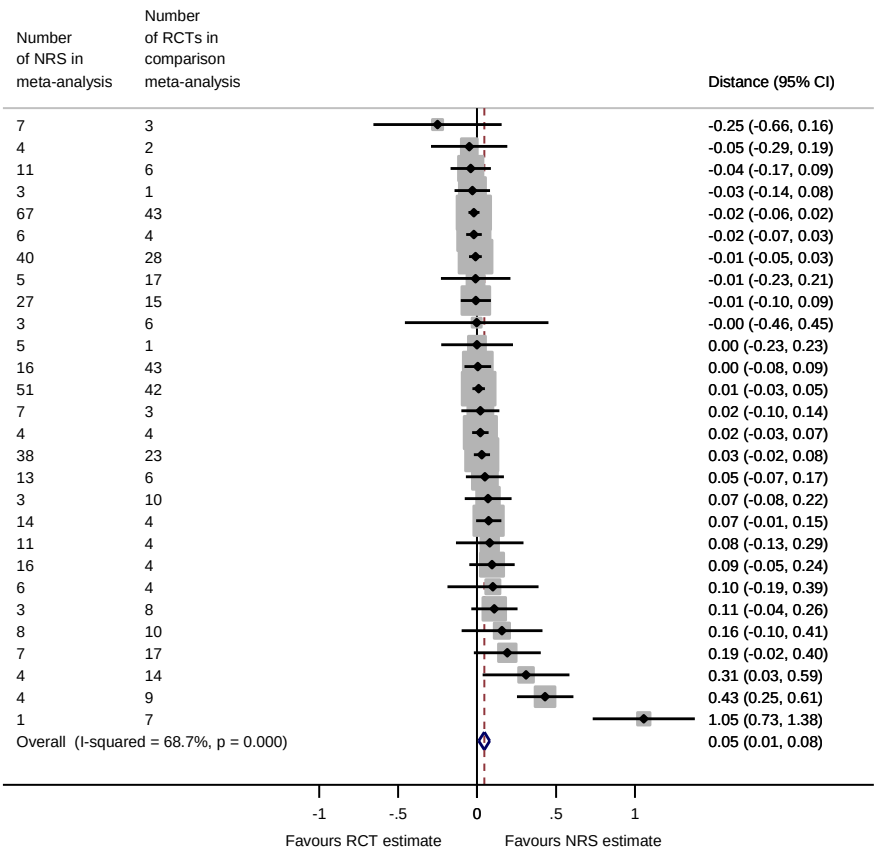


Figure 5.2 Number of pooled effects by risk of bias



Pooled effects were estimated using inverse-variance weighted random effects meta-analysis. A random effects model was used under the assumption that bias (and different degrees of bias), may differentially affect the estimates across topics; for example, confounding due to self-selection into the intervention may be thought more likely to affect interventions targeting individuals, like microcredit or entrepreneurship training, than those targeting groups, such as most education interventions. The analysis (Figure 5.3) suggests that, on average, NRS estimated a slightly bigger pooled effect than RCTs for the same pair-wise intervention-outcome (D=0.05, 95%CI=0.01, 0.08; number of pair-wise meta-analysis comparisons=28). There is estimated statistical heterogeneity in the distribution of pooled effects (I-squared=69%; tau-squared=0.004) and the within-meta-analysis variance of each pooled effect is inversely proportional to the number of studies contained in the meta-analysis (correlation=0.53, p<0.005).

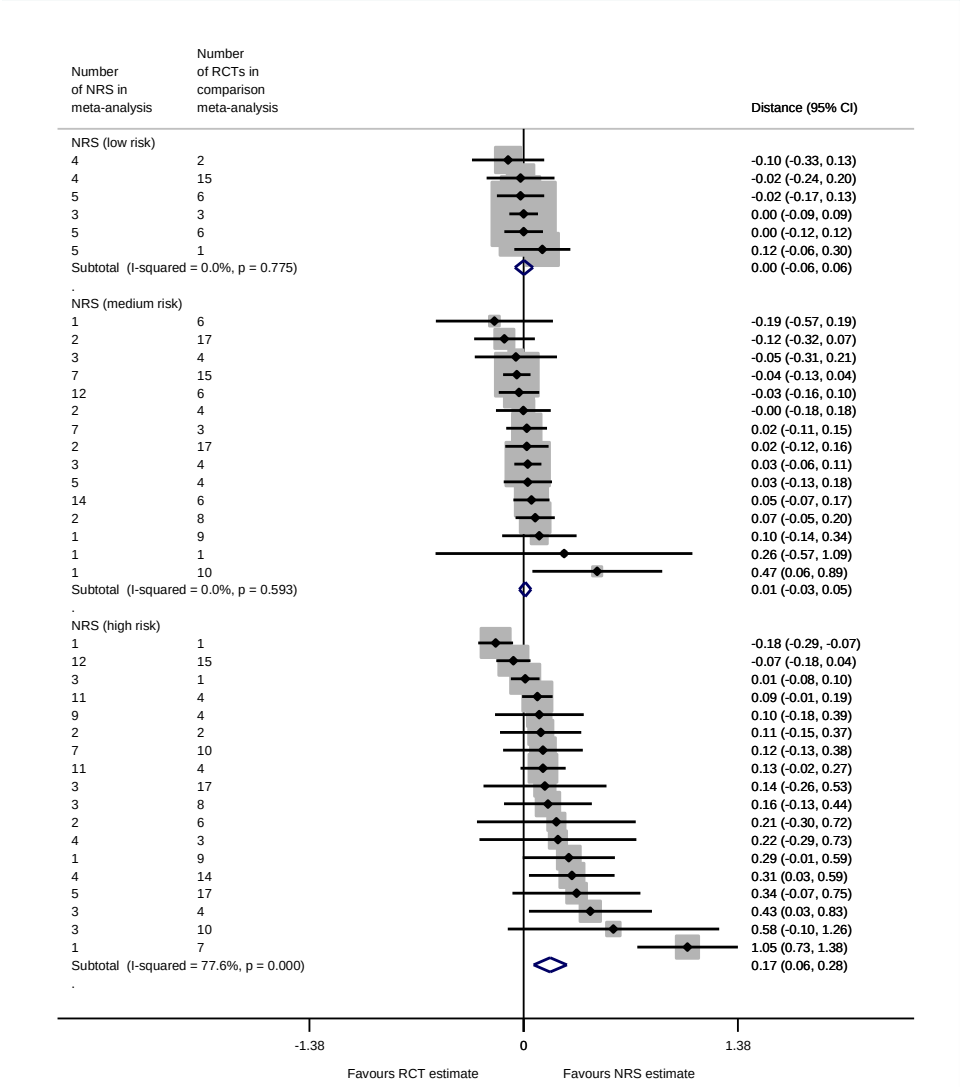
Figure 5.3 Meta-analyses comparing NRS and RCTs



Analysis was therefore done to explore whether greater probability of bias in underlying studies was associated with greater deviation of the pooled effect.

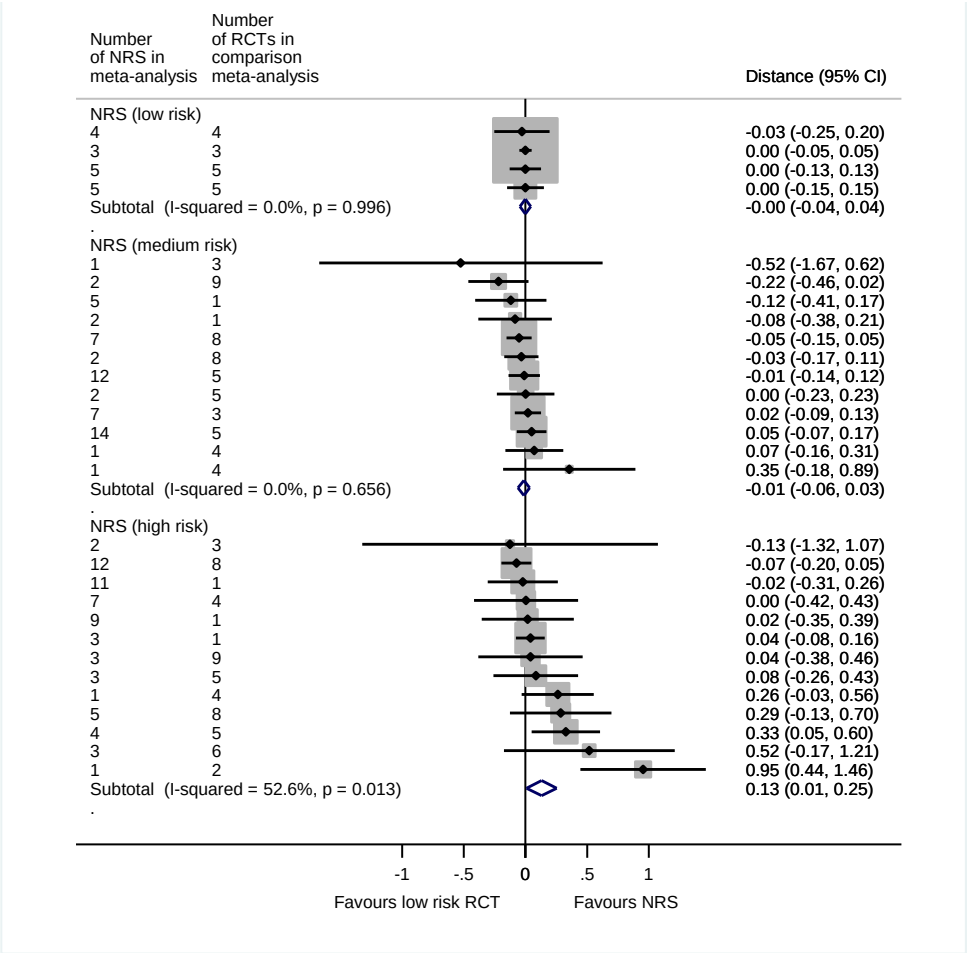
The findings indicate that, for NRS on average, for which the pair-wise comparison was the pooled effect across RCTs, bias was positively correlated with distance (Figure 5.4).

Figure 5.4 Meta-analyses by NRS risk of bias



Additional meta-analyses were estimated for 'low risk of bias' comparators (Figure 5.5 and Figure 5.6). Table 5.3 presents the summary findings, comparing 'low risk of bias' RCTs with NRS. On average, the biggest deviation from RCTs was given by NRS with 'high risk of bias' (D=0.17, 95%CI=0.07, 0.28; 18 pair-wise meta-analysis comparisons), whereas NRS with 'low risk of bias' produced the same effects as RCTs on average (D=0.00, 95%CI=-0.06, 0.06; 6 pair-wise meta-analyses) with no residual heterogeneity (I-squared=0%; tau-squared=0).

Figure 5.5 Meta-analyses of NRS versus low-risk RCTs



In contrast, distance is inversely correlated with probability of bias for RCTs, on average (Figure 5.6). Thus, for example, pooled effects from meta-analyses of RCTs assessed of being of 'high risk of bias' were on average 0.08 standard deviations smaller than effects from meta-analyses of RCTs of 'low risk of bias' (95%CI=-0.14, -0.03; 10 pair-wise meta-analyses) with zero estimated statistical heterogeneity measured relatively as a percentage of total variation (I-squared=0%) or absolutely in the units of the effect size (tau-squared=0.000). It is not clear why this might be the case – that, in contrast to findings for NRS, meta-analyses of RCTs with greater probability of bias produce smaller effects on average than those of RCTs with low probability of bias – although one possible explanation is that the analyses are confounded by unobserved heterogeneity. Possibly, RCTs estimating bigger effects are better designed and implemented, including better fidelity of intervention, and/or are subject to evaluation placement bias, also called site-selection bias (Allcott, 2015).

Figure 5.6 Meta-analyses of RCTs versus low-risk RCTs

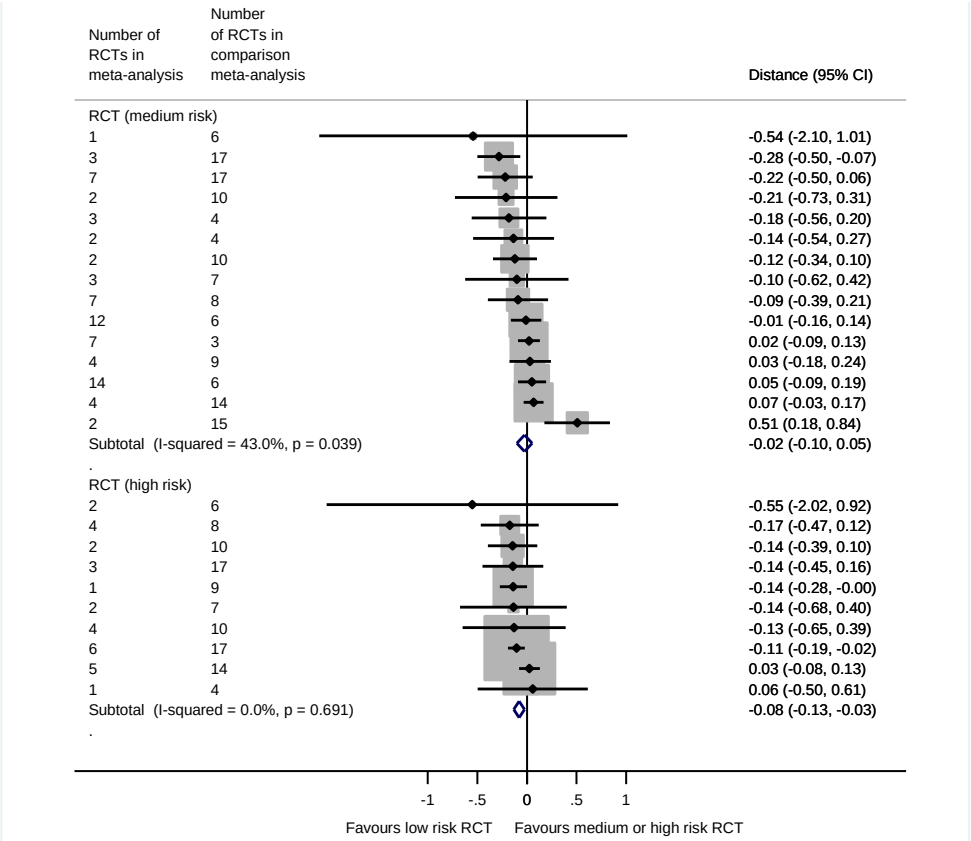


Table 5.3 Random effects meta-analysis of distance statistics

Comparison	*D*	*95% confidence interval*		*I²*	*Tau²*	*N*
NRS – RCT	**0.045**	**0.010**	**0.080**	68%	0.004	28
NRS (low) – RCT	0.002	-0.056	0.060	0%	0.000	6
NRS (some) – RCT	0.010	-0.027	0.048	0%	0.000	15
NRS (high) – RCT	**0.171**	**0.065**	**0.278**	78%	0.033	18
RCT (some) – RCT (low)	-0.024	-0.102	0.053	43%	0.008	15
RCT (high) – RCT (low)	**-0.080**	**-0.135**	**-0.026**	0%	0.000	10
NRS (low) – RCT (low)	-0.001	-0.044	0.042	0%	0.000	4
NRS (some) – RCT (low)	-0.013	-0.060	0.034	0%	0.000	12
NRS (high) – RCT (low)	**0.130**	**0.008**	**0.253**	53%	0.021	13

Notes: **bold** indicates *D* is statistically significantly different from zero at less than 5% significance; low, some, high refer to 'low risk of bias', 'some concerns' and 'high risk of bias', respectively.

Finally, sensitivity analysis was performed by excluding pooled effects from analysis where the number of RCTs or NRS was small (Table 5.4). Distance estimates were of smaller magnitude when 'pooled effects' containing only a single study (whether RCT or non-randomised study) were excluded from meta-analysis, but the signs and statistical significance were consistent with previous findings.[107]

Table 5.4 Random effects meta-analysis excluding small sample sizes

Comparison	*D*	*95% confidence interval*		I^2	Tau^2	*N*
NRS – RCT	**0.029**	**0.002**	**0.056**	48%	0.002	25
NRS (low) – RCT	-0.012	-0.073	0.049	0%	0.000	5
NRS (some) – RCT	0.006	-0.033	0.044	0%	0.000	11
NRS (high) – RCT	**0.117**	**0.042**	**0.184**	25%	0.004	14
RCT (some) – RCT (low)	-0.024	-0.102	0.055	46%	0.009	14
RCT (high) – RCT (low)	**-0.071**	**-0.131**	**-0.011**	0%	0.000	8
NRS (low) – RCT (low)	-0.001	-0.044	0.042	0%	0.000	4
NRS (some) – RCT (low)	-0.019	-0.067	0.030	0%	0.000	9
NRS (high) – RCT (low)	0.057	-0.055	0.169	17%	0.005	10

Notes: **bold** indicates *D* is statistically significantly different from zero at less than 5% significance; low, some, high refer to 'low risk of bias', 'some concerns' and 'high risk of bias', respectively.

It is usually argued in systematic reviews that NRS are included for ecological validity (i.e., relevance to the 'real world' of intervention programming) or other factors relating to external validity, such as measuring longer-term consequences (e.g., Welch et al., 2016). This analysis also suggests that well-designed and implemented NRS provide internally valid effect estimates. The included meta-analyses cover a range of topic areas and geographies, suggesting the findings are generalisable across interventions, outcomes and contexts.

[107] Sensitivity analysis excluding pooled effects from fewer than three studies produced the same findings in magnitudes, signs and statistical significance of distance estimates.

A factor that may be systematically correlated with effect sizes across intervention-outcome pair-wise comparisons of RCTs and NRS, is the external validity of the estimate (see Chapter 4, Section 4.5). Cross-study comparisons, such as those presented in this section, compare studies conducted among different underlying populations. There are concerns about the validity of these comparisons in quantifying bias, even when these studies find zero differences in treatment effects across RCTs and NRS (or different degrees of bias) on average. Even when the non-randomised study generates an unbiased treatment effect estimate for the sample, there may still be a difference in effect with a comparable, well-conducted randomised study because of: 1) sampling error, which would tend to zero in expectation in meta-analysis; and 2) sampling bias due to the characteristics of the sample included in analysis – for example, the average treatment effect (ATE) causal estimand from an RCT, the local average treatment effect (LATE) estimand in RDD, the average treatment effect on the treated (ATET) in statistical matching and double differences, or the complier average causal effect (CACE) in instrumental variables estimation. Cook et al. (2008) stated that there is no theoretical reason why one should expect these differences to 'cancel out' on average in meta-analysis.

However, it is not clear why, if the systematic difference in effect sizes between NRS and RCTs on average were related to external validity, that difference would only be apparent for higher risk-of-bias NRS and not all NRS regardless of bias probability. Indeed, when pair-wise comparisons are made between NRS and RCTs with 'low risk of bias', the difference in mean pooled effects is only significant for NRS with 'high risk of bias' (Table 5.3). Even so, it is not possible to rule out the possibility of the apparent difference by risk-of-bias status being confounded by external validity. Therefore, Section 5.3 considers this potential source of systematic variation, by analysing data from a systematic review of internal replication studies in international development.

5.3 Systematic review of within-study comparisons in international development

The conceptually preferred approach to empirical measurement of bias is the 'internal replication study' (Cook et al., 2008) or 'design replications study' (Wong and Steiner, 2016). Like cross-study comparisons, these compare a particular estimator, usually a non-randomised comparison group, with a causal benchmark, usually an RCT, which is assumed to provide an unbiased estimate. However, the comparison arm used in the NRS comes from the same target population, hence they are also called 'within study comparisons' (Bloom et al., 2002; Glazerman et al., 2003). They have been conducted in the social sciences since the 1980s, following an internal replication of the randomised evaluation of the National Supported Work (NSW) Demonstration programme in the U.S.A. (Lalonde, 1986), and a large number of reviews of these studies exists (Appendix B).

Glazerman et al. (2003, p.65) defined an internal replication study as follows: "researchers estimate a program's impact by using a randomized control group and then re-estimate the impact by using one or more non-randomized comparison groups." There are four main ways of doing internal replication studies, involving varying degrees of data requirements (Wong and Steiner, 2016). The most data intensive is called independent design, or 'four-arm' design (Shadish et al., 2008). Participants are randomly assigned into RCT and NRS arms. Subsequently, participants in the RCT arm are randomly assigned into treatment and control, and those in the NRS arm self-select or are selected by a third party into a preferred treatment option. The difference in the estimated treatment effects in RCT and NRS is then calculated, to form the bias estimate, from the four independent arms.

In contrast, all other internal replication designs have some degree of dependency across arms; usually, the RCT treatment arm is common across study arms, and a non-equivalent comparison group is created, which is compared to the RCT control group mean. In 'simultaneous design', observations drawn from an overall population are selected to participate in the RCT. The corresponding NRS uses administrative data or an observational study from a sample of the target population that did not participate in the RCT (e.g., Lalonde, 1986). However, the assumption of the design is that measurement of the same outcome at the same time, under the same study conditions, in NRS and RCT – that is, comparability of the target

population – factors which are difficult to satisfy in practice (Smith and Todd, 2005). 'Multi-site simultaneous design' attempts to account for this by using data from an RCT based on multiple selected sites, within each of which participants are randomly assigned to treatment and control. The NRS is constructed by comparing average outcomes from the RCT treatment group in one site to the control observations from another site. Another type of simultaneous design, called a 'tie-breaker' design by Chaplin et al. (2018), is done to enable comparison of the cluster-RCT and RDD estimators. The initial selection of clusters into the benchmark study is determined by an eligibility criterion, usually a threshold score, after which random assignment is done. The NRS compares observations within clusters immediately around the eligibility threshold – control observations from the RCT with comparison observations on the other side of the threshold which were not eligible for the RCT. Group eligibility for several conditional cash transfer (CCT) programmes was assessed this way, therefore these programmes feature heavily in RDD within-study comparisons in international development (e.g., Buddelmeyer and Skoufias, 2005).

Finally, in 'synthetic design', which is the least data intensive, the researcher simulates the NRS from existing RCT data by removing observations from the treatment and/or control arm to create non-equivalent groups. For example, Fretheim et al. (2013) discarded control group data from a cluster-RCT with 12 months of outcome data points available from health administrative records before and after intervention, in order to compare the findings with interrupted time series analysis.[108] Synthetic design is also used to assess validity of RDD in cluster-RCTs where pre-test discriminant score data are available to compare participants from eligible clusters into 'treatment' and those from ineligible clusters forming the 'comparison' arm of the RDD (Wong and Steiner, 2016).[109] Hence, the main difference between

[108] In further analysis of additional studies, Fretheim et al. (2015) discarded control group data from four cluster-RCTs of medical interventions containing six or more time series data points for outcomes before and after intervention, in order to compare the findings with interrupted time series. The authors also incorporated control group observations compare the findings with controlled interrupted time series analysis. An interesting finding of the study, which also included ITS with between and five pre-intervention periods, was that the findings of ITS were less reliable than those with at least six pre-intervention periods.

[109] This approach was also used in the group A (eligible households in treated clusters) versus group D (ineligible households in control clusters) comparisons in Buddelmeyer and Skoufias (2005), and in the 'pure control' group comparisons in Barrera-Osorio et al. (2014).

simultaneous and synthetic design is that, in the latter, the researcher removes observations to exploit a single dimension of variation.

5.3.1 *Existing reviews of within-study comparisons*

Evidence from internal replication studies suggests that, when inappropriately designed or executed, NRS are likely to yield effect size estimates that do not statistically correspond to RCTs; a factor which is, sometimes inappropriately, assumed to represent bias. The first meta-analysis of evidence from internal replication studies synthesised 12 evaluations of employment programmes on earnings (Glazerman et al., 2002, 2003). All studies originated in high income contexts and three-quarters of the interventions and data collection were undertaken in the 1970s and 80s.[110] The analysis is of the correspondence between RCT and NRS findings from a range of different NRS approaches (including cross-section regression, fixed effects panel and double differences regression, statistical matching and selection models). It concluded that NRS rarely replicated experimental estimates and the absolute magnitude of the differences was often quite large, equivalent to 10 percent of annual earnings in some instances. However, the extent to which evidence of statistical correspondence with RCT estimates adequately represents bias in NRS findings depends on quality of implementation of RCT and NRS (internal validity), and possibilities for confounding by differences in the target population (external validity).

Regarding internal validity, Cook et al. (2008) showed that NRS in which the method of treatment assignment is known or carefully modelled using baseline data, produced very similar findings in direct comparisons with RCTs. Hansen et al. (2013) made the first review of evidence from development interventions, including internal replication studies of two cluster-RCTs of CCT programmes (*Programa de Educación, Salud y Alimentación*, PROGRESA, in Mexico and *Red de Protección Social*, RPS, in Nicaragua), and an individually randomised lottery of migration visas in Tonga. One replication examined the correspondence of estimates from regression discontinuity design (Buddelmeyer and Skoufias, 2004), and the others examined the correspondence of double differences, statistical matching and instrumental variables estimation (Diaz and Handa, 2006;

[110] Four studies addressed the same intervention, the U.S. NSW Demonstration.

Handa and Maluccio, 2010; McKenzie et al., 2010). The review found that the difference between NRS estimates and RCTs was smaller where self-selection into treatment was more negligible and the selection process simple or well understood.

One of the NRS approaches commonly thought to produce internally valid estimators in expectation is the regression discontinuity design (e.g., Shadish et al., 2002). Chaplin et al. (2018) assessed the statistical correspondence of 15 internal replication studies with an RDD approach (including two studies based on data collected on programmes in L&MICs) using meta-analysis. The average distance between RCT and RDD estimates was approximately 0.1 standard deviations. However, they warned that researchers should not assume based on these findings that individual RDD estimates would necessarily be near zero, suggesting factors such as larger samples and the choice of bandwidth may prove important in determining the degree of bias in an individual RDD estimate.

However, Smith and Todd (2005) warned against "searching for 'the' nonexperimental estimator that will always solve the selection bias problem inherent in nonexperimental evaluations" (p.306). Instead, they argued research should seek to map and understand the contexts that may influence studies' degrees of bias. For instance, Hansen et al. (2013) noted the potential importance of the type of dependent variable examined in studies, suggesting simple variables (such as binary indicators of school attendance) may be easier to model relative to more complex outcome variables (such as consumption expenditure or earnings), although presumably this could also relate to the use of observed rather than self-reported outcomes. Glazerman et al. (2003) found that the data source, the quality of control variables, and evidence of statistical robustness tests, were related to the magnitude of estimator bias. Synthesising results from 12 internal replication studies (all from high income countries) of standardised reading or math test scores, Wong et al. (2017) found that use of baseline outcomes, geographical proximity of treatment and comparison, and breadth of control variables, were associated with smaller distance between RCT and NRS. They also noted that NRS that simply relied on a set of demographic variables, or prioritised local matching when local comparisons were not comparable to treated cases, rarely replicated RCT estimates.

The second main potential source of discrepancy between the findings of RCTs and NRS is in the effect size quantity or estimand due to differences in the target population in each study (external validity) (e.g., Duvendack et al., 2012), also called sampling bias. For example, confounding may occur when attempting to compare an average treatment effect (ATE) estimate from an RCT with ATET from a double difference or matching study, or LATE from an RDD (Cook et al., 2008). The ITT estimator, on which ATE is based on RCTs, becomes smaller as non-adherence increases, making raw comparison of the two estimators inappropriate, even if they are both unbiased. Similarly, LATE may be an unbiased estimate of the average effect of an intervention amongst the population immediately around the treatment threshold in RDD, but it may still differ from ATE where the treatment effect is not constant across the sample or population receiving treatment.

Cook et al. (2008) stated that, due to the potential for results-based choices in the covariates and methods used – called 'specification searches' (Leamer, 1978, 1983) – NRS analysts should be blinded to the results of the RCT they are replicating. Hansen et al. (2010) note that where the benchmark result is known, any findings illustrate "that a comparable estimate *can be* found, not that it *will be* found in practice" (p.331; original emphasis by authors). These biases may serve to accentuate or diminish the differences between RCT and NRS depending on the replication study authors' priors. Thus, Freitheim et al. (2015) "concealed the results and discussion sections in the retrieved articles using 3M Post-it notes and attempted to remain blinded to the original results until after our analyses had been completed" (p.326). Where it is not possible to blind replication researchers to the RCT findings, which would usually be the case, a reasonable expectation is that the internal replication report should contain sensitivity analysis documenting differences in effects due to changes in specification (Hansen et al., 2013). In addition, a distinct advantage of the latter approach, whether done openly or blinded, is to enable the assessment of sensitivity analysis to different methods of implementation in the particular NRS (e.g., matching at group or individual level, inclusion of baseline outcome, the importance of using geographically proximate observations).

Most existing reviews of internal replication studies have not been done systematically – that is, based on systematic approaches to identify and critically appraise studies and statistically synthesise effect size findings.

Exceptions include a review by Wong et al. (2017), which reported a systematic search strategy, and Glazerman et al. (2003) and Chaplin et al. (2018), which used statistical meta-analysis of effect sizes. The existing review of evidence from social and economic development programmes in L&MICs (Hansen et al., 2013) did not report systematic methods of search, critical appraisal of included benchmark and replication studies, or effect size analysis. This study was therefore updated to incorporate more recent internal replication studies and methods of analysis – that is, 'update search' and 'update quality' (Waddington et al., 2018).

5.3.2 *Study inclusion decisions*

The eligibility criteria for inclusion in the review are given in Table 5.5. All included studies reported treatment effects for a causal benchmark study (a sample randomly assigned in an experimental or natural experimental context), and for a non-randomly assigned comparison replication. Eligible benchmark studies needed to use randomised assignment, whether controlled by researchers (RCTs) or policymakers (randomised natural experiment). Eligible within-study comparisons included any non-randomised approach, whether natural experiment, quasi-experiment or pure observational study with selection on observables. These included methods with adjustment for unobservable confounding, such as DD, IV, RDD and methods adjusting for observables only such as statistical matching and OLS. A rationale for excluding OLS is that, unlike matching, it cannot account for biases arising from comparing observationally dissimilar groups (Heckman et al., 1997); however, it does estimate the treatment estimand over the same target population as the randomised benchmark (average treatment effect, ATE).

An important criterion for inclusion was that the NRS and benchmark estimated the same treatment estimand, or equivalently, where the bias estimator used the benchmark control and NRS comparison means only, data were from the same target population. As discussed, this is important to avoid confounding of the bias estimator. Evidence suggests that the assumption of constant treatment effects across samples, which would be necessary to validate comparison of different treatment estimands, should not be relied on. For example, Oosterbeek et al. (2008) showed a positive impact on school enrolment for the poorest quintile receiving benefits under

the *Bono de Desarrollo Humano* (BDH) CCT programme in Ecuador, but no impact for the second poorest quintile.

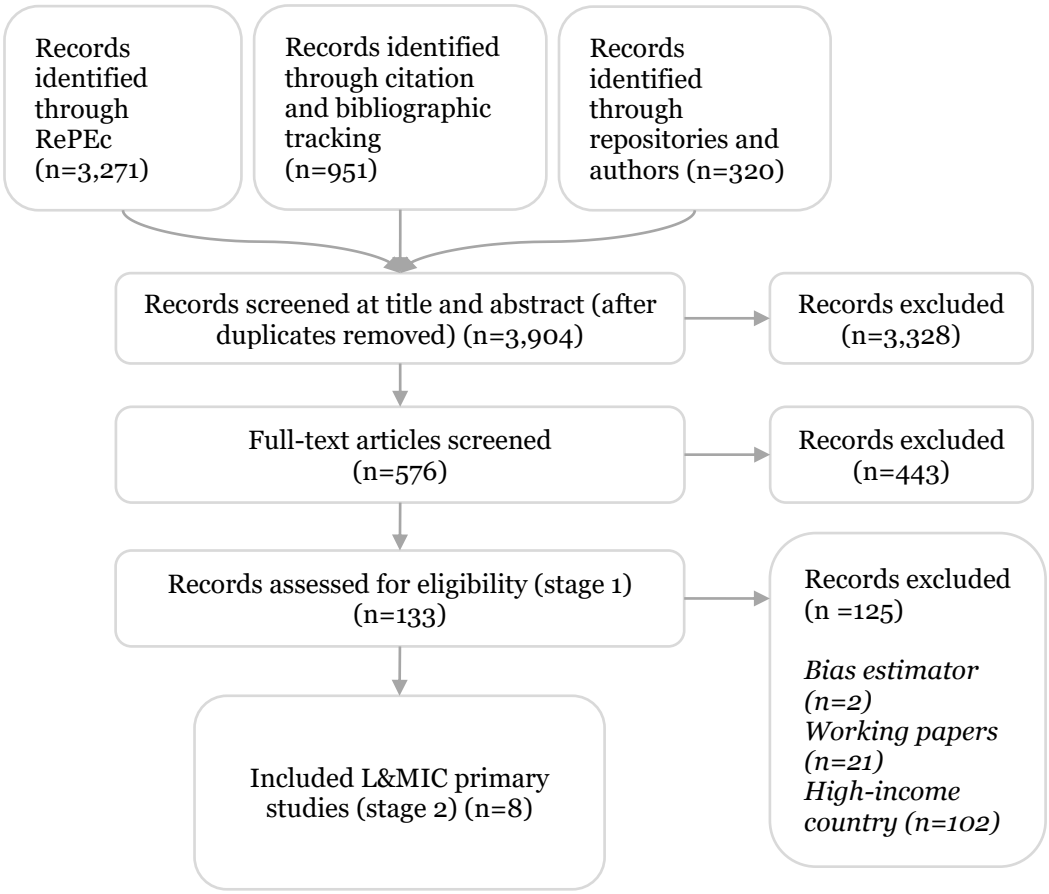
Table 5.5 Inclusion criteria of review of internal replication studies

Criteria	*Included studies*	*Excluded studies*
Population	General programme participants in L&MICs.	Programmes conducted among populations in high income country contexts (e.g., Fretheim et al., 2013).
Intervention and comparator	Any social or economic development intervention requiring behaviour change and any comparison condition (e.g., no intervention, wait-list, alternate intervention).	Clinical or bio-medical interventions.
Benchmark study design	Within-study comparisons reporting results of a benchmark randomised study, where randomisation was done by researchers or a public lottery.	Within-study comparisons where the causal benchmark was not randomly assigned (e.g., Friedman et al., 2016).
NRS study design	Within-study comparisons reporting results of NRS comparison replication using any method (e.g., DD, IV, OLS, Matching, RDD) from same target population and time period, using the same outcome as benchmark study.	Within-study comparisons where there is no overlap in treatment group samples for benchmark and comparison (e.g., Glewwe et al., 2004), or where target population differs (e.g., Oosterbeek et al., 2008; Urquieta et al., 2009; Lamadrid-Figueroa et al., 2013).

Previous reviews noted several issues in systematically identifying internal replication studies due to a lack of common language used to systematically index this evidence. Glazerman et al. (2003) indicated electronic searches failed to comprehensively identify many known studies, while Chaplin et al. (2018) stated that, despite attempting to search broadly, "we cannot even be sure of having found all past relevant studies" (p.424). Hence, a combination of search methods was used, including: electronic searches of databases, where search terms were identified using 'pearl harvesting' (using keywords from known eligible studies) (Sandieson, 2006); bibliographic back-referencing of bibliographies of included studies and reviews of internal replication studies; forward citation tracing of reviews of internal replication studies using three electronic tracking systems (Google Scholar, Web of Science and Scopus); hand searches of the repository of a known institutional provider of internal replication studies (Manpower Demonstration Research Corporation, MDRC); and by contacting authors. Full details of the search

strategy are in Appendix B. Following identification of 3,904 non-duplicate records, 133 were assessed as being eligible at stage 1 (study conducted in any geographic location) and finally eight studies were assessed as eligible for the review at stage 2 (L&MIC location only) (Figure 5.7).

Figure 5.7 PRISMA flow diagram for internal replication studies



Source: Villar and Waddington (2019).

A number of studies made comparisons between randomised and non-randomised estimates of programmes among L&MIC populations, but nevertheless were excluded from review. For example, Friedman et al. (2016) was excluded because the study examined the difference between IV and OLS estimation for the effect of education attainment on outcomes, rather than the effect of the randomised scholarship programme. Cintina and Love (2014) created non-randomised treatment and control groups from the microcredit RCT in India by Banerjee et al. (2013), aiming to answer research questions relating to relative effectiveness of interventions and spillover effects (similar to Angelucci and de Giorgi, 2006, and Barrientos and

Sabates-Wheeler, 2011),[111] and as such, did not provide an estimate of effect of the same intervention using randomised and non-randomised groups. Finally, Miguel and Kremer (2003) presented RCT and 'nonexperimental' estimates of the relationship between child deworming adherence and various indicators of social networks in Kenya, but it appears that the 'nonexperimental' estimators presented by the authors are simply average treatment effect on the treated (ATET) estimates from the RCT sample, rather than estimates using another source of data for the NRS comparison.

Several studies were excluded due to differences in target population and/or intervention receipt between RCT and NRS. Oosterbeek et al. (2008) compared the findings of a randomised experiment of the BDH programme in Ecuador, conducted among households with a poverty index value between the 13th and 28th poverty percentiles, with an RDD analysis among households between the 33rd and the 47th percentiles. Two other RDD internal replication studies of the *Oportunidades* CCT scheme in Mexico, were excluded because they did not present local average treatment effect results from the randomised benchmark within the RDD bandwidth, and hence the findings were confounded by causal estimand (Urquieta et al., 2009; Lamadrid-Figueroa et al., 2013). Behrman et al. (2009) compared randomised and non-randomised estimates using a matched comparison measured at a different point in time with differential exposure to *Oportunidades*. Barham et al. (2014) compared randomised and non-randomised estimates from different years for RPS in Nicaragua. Glewwe et al. (2004) examined differences between different interventions and target populations of education programmes in Kenya, therefore undertaking between-study comparison.

Eight eligible internal replications were included of randomised studies of social and economic programmes in L&MICs (Table 5.6). Four of these featured in the previous review of internal replication studies in international development (Hansen et al., 2013). These were based on data from two CCT programmes, PROGRESA in Mexico (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006) and RPS in Nicaragua (Handa and Maluccio, 2010), plus

[111] Gertler et al. (2012) also replicated RCT findings for the *Oportunidades* CCT programme in Mexico, in order to test for general equilibrium effects (a form of spillover that affects the underlying incentives that operate in a local economy or more widely, such as prices and wages). However, the replication uses ineligible participants from the RCT data without estimating a NRS comparison.

a randomised lottery balloting permanent migration visas in Tonga (McKenzie et al., 2010).[112] One study on PROGRESA examined the correspondence of estimates from RDD analysis with estimates from an RCT (Buddelmeyer and Skoufias, 2004). An additional four replications of RCTs were located through the searches, including two of the *Programa de Asignación Familiar* (PRAF) in Honduras (Galiani and McEwan, 2013; Galiani et al., 2017), one of a scholarship programme in Cambodia (Barrera-Osorio et al., 2014) and one of electricity subsidies in Tanzania (Chaplin et al., 2017).

All but one included study used a randomised field trial (RCT) as the benchmark. McKenzie et al. (2010) used a randomised natural experiment, where programme assignment was done by a public lottery by policy makers, and the data itself were collected by the authors specifically to estimate the treatment effect of the lottery. Clusters were randomly assigned to the programme in Galiani and McEwan (2013) and Galiani et al. (2017) as part of a field trial, but the study used administrative data to evaluate outcomes (from the national census), hence the studies have the benefits of blinding of outcomes data collection because reporting is not linked to programme participation by participants or outcome assessors.[113]

[112] The visas enabled Tongans to enjoy permanent residency in New Zealand under the PAC (New Zealand's immigration policy which allows an annual quota of Tongans to migrate).

[113] In this sense, according to the classification approach presented in Chapter 4, Galiani and McEwan (2013) and Galiani et al. (2017) are true natural experiments whereas McKenzie et al. (2010) is a randomised quasi-experiment.

Table 5.6 Eligible within-study comparisons of development programmes

Study	*Intervention*	*Country*	*Outcome(s)*	*Benchmark*	*NRS replication*	*WSC type*
Buddelmeyer and Skoufias (2004)	Cash transfer (PROGRESA)	Mexico	Reported school attendance and child labour	Cluster-RCT	RDD	Tiebreaker
Diaz and Handa (2006)	Cash transfer (PROGRESA)	Mexico	Reported food expenditure, school enrolment, child labour	Cluster-RCT	OLS, matching	Simultaneous
Handa and Maluccio (2010)	Cash transfer (RPS)	Nicaragua	Reported expenditure, childcare, preventive health care, child illness	Cluster-RCT	Matching	Simultaneous
McKenzie et al. (2010)	Immigration entitlement	Tonga	Reported income	Randomised natural experiment	DD, IV, OLS, Matching	Simultaneous
Galiani and McEwan (2013)	Cash transfer (PRAF)	Honduras	Census reported school enrolment and child labour	Cluster-RCT	RDD	Tiebreaker
Barrera-Osorio et al. (2014)	Scholarship	Cambodia	Grade completion and math test score	Cluster-RCT	RDD	Tiebreaker
Chaplin et al. (2017)	Subsidy	Tanzania	Reported energy use and cost	Cluster-RCT	Matching	Simultaneous
Galiani et al. (2017)	Cash Transfer (PRAF)	Honduras	Census reported school enrolment and child labour	Cluster-RCT	GDD	Tiebreaker

The studies tested a range of non-randomised replication methods including geographical discontinuity design (GDD),[114] RDD, IV, PSM, and DD, all using variants of simultaneous design. All discontinuity design replications included were able to restrict the RCT samples to create localised randomised estimates in the vicinity of the discontinuity (local average treatment effects), and compared the distance between the two treatment effect estimates (Buddelmeyer and Skoufias, 2004; Galiani and McEwan, 2013; Barrera-Osorio et al., 2014; Galiani et al., 2017). In the case of Galiani and McEwan (2013), programme eligibility was set for localities below a threshold on mean height-for-age z-score of -2.304. The benchmark sample was therefore restricted to the block of localities with mean z-score just below the threshold. The RDD comparison was generated for untreated localities just above the threshold, where the z-score was predicted for comparisons due to limited data. In Buddelmeyer and Skoufias (2005), there were four groups of households which enabled the RDD estimator to be compared to the RCT. The groups were differentiated by treatment status of the cluster, determined by randomisation across those clusters below a maximum discriminant score (poverty index); and eligibility of households within clusters for treatment, determined by the household's discriminant score.[115] The RCT treatment estimand was calculated over households within the same bandwidth as the RDDs in order to ensure comparability of the target population.

Other studies used statistical methods to compare NRS comparison groups with randomised control group means (Diaz and Handa, 2006; Handa and Maluccio, 2010; McKenzie et al., 2010; Chaplin et al., 2017).

[114] Galiani et al. (2017) stated that it was unlikely that households from the indigenous Lenca group migrated to obtain benefits under the CCT programme, suggesting validity of the benchmark control group. However, there remained differences in shares of Lenca populations across the geographical discontinuity in cash transfer treatment and control communities, potentially invalidating the GDD comparison. Therefore, the study design should be considered a 'geographical quasi-experiment' where potential outcomes are assumed independent of treatment assignment, conditional on observed covariates.

[115] The comparisons used in RDD were: group A (eligible households in treated communities) versus group C (ineligible households in treated communities); group A versus group D (ineligible households in control communities); group C versus group B (eligible households in control communities); and group C versus group D (ineligible households in treatment and control communities). The final comparison, group A versus group B (eligible households in control communities) was used to calculate the RCT treatment effect, which. The replication researchers therefore appear to use 'synthetic design' for the A versus D comparison (for definitions, see Wong and Steiner, 2016).

5.3.3 *Risk of bias in within-study comparison estimate*

Existing reviews of internal replication studies do not provide comprehensive assessments of the risk of bias to the effect estimate in the benchmark study using formal risk-of-bias tools. Partial exceptions are Glazerman et al. (2003), who commented on the likely validity of the benchmark RCTs (randomisation oversight, performance bias and attrition), and Chaplin et al. (2018) who coded information on use of covariates to control for pre-existing differences across groups and use of balance tests in estimation.

Modified applications of Cochrane's tools for assessing risk of bias in RCTs were used to assess biases in benchmark cluster-randomised (Eldridge et al., 2016) and individually randomised studies (Higgins et al., 2016).[116] For the benchmark (individually randomised) natural experiment (McKenzie et al., 2010), which was analysed using instrumental variables due to non-compliance, the risk-of-bias assessment drew on the tool developed in this Thesis (Appendix A Table A2) as well as relevant questions about selection bias into the study from Eldridge et al. (2016).[117] In addition, the appraisal of the benchmark took into account the relevance of the bias domains for determining relative bias between NRS and randomised estimators was also used in the bias assessment, as well as factors that may have confounded the estimated difference between benchmark and NRS replication in estimation of the with-study comparison.

Only one randomised benchmark had 'low risk of bias' (Galiani and McEwan, 2013; Galiani et al., 2017). However, due to problems in implementing the NRS in those studies, there remained 'some concerns' about confounding of the NRS-RCT distance estimate with respect to its interpretation as bias. The benchmark for PROGRESA had 'high risk of bias' (Buddelmeyer and Skoufias, 2004). The remaining benchmark studies had 'some concerns', and

[116] It was not considered necessary to blind coders to results following Cook et al. (2008) – for example, by removing the numeric results and the descriptions of results (including relevant text from abstract and conclusion), as well as any identifying items such as author's names, study titles, year of study and details of publication – since all studies reported multiple within-study comparisons and all data were extracted and analysed by the author.

[117] Cochrane's risk of bias tool for RCTs does not enable the reviewer to discern the validity of the application of IV to correct for non-compliance. The maximum score available in that tool under non-compliance, even under appropriately conducted IV, is 'some concerns'.

there were a few instances 'high risk of bias' in the NRS replications due to differences in definition of outcomes with the benchmark survey questions (Diaz and Handa, 2006; Handa and Maluccio, 2010). Hence, all the within-study comparison estimates of bias may be confounded (Table 5.6).

Concerns about the benchmarks often arose from a lack of information, such as in the case of attrition in the PROGRESA benchmark experiment, or in assessing imbalance of baseline characteristics using distance metrics. In other instances, concerns were more difficult to address. For example, none of the studies were able to blind participants or outcome assessors to intervention, while outcomes were mainly collected through self-report. For benchmark studies using cluster-randomisation, where informed consent does not alert participants to the intervention (and outcome assessors may also be blinded), this source of bias may be less problematic (Schmidt, 2014). And with respect to recruitment of participants and deviation from intended interventions, it is not clear that evaluations of social programmes administered to clusters, where participants are identified after cluster assignment, as in the case of PROGRESA, can sufficiently capture data non-adherence due to participant migration between clusters.

However, it was not always clear that the risk of absolute bias arising in the benchmark estimate, would necessarily lead to a difference in relative bias in the difference estimate. For example, threats to validity due to incomplete treatment implementation under 'deviations from intended intervention' are not necessarily threats to validity in the distance estimate for the within-study comparison, which is made by comparing the randomised control and NRS comparison means. Similarly, absolute biases arising due to collection of reported outcomes data in open trials may not cause relative bias if the NRS uses the same data collection methods, and the potential sources of bias in benchmark and observational study are considered to be equivalent. Bias due to selective reporting that may have affected benchmark trials was not judged problematic in the context of within-study comparisons, where multiple specifications, outcomes and sub-groups were often included to provide diversity in the estimates.

Finally, relative bias in the difference estimate may be caused by bias in the NRS and confounding of the relationship. Bias in the NRS is captured in the analysis of distance estimates, in order to inform the risk-of-bias tool.

Confounding of the relationship may occur primarily for two reasons – due to differences in the survey instrument (e.g., outcome measurement) and target population. The rest of this section discusses the critical appraisal domains in turn that relate to absolute bias in the benchmark estimate (Sections 5.3.3.1 to 5.3.3.6), as well as any factors that are relevant in assessing whether the bias also applies to relative bias between benchmark and NRS (Section 5.3.3.7).

5.3.3.1 Confounding

Benchmark study data are typically from cluster-randomised field trials, five of which evaluated conditional cash transfers in Latin America. These programmes were typically randomised at public events with members of the government, media and field research teams present. Two benchmarks were assessed as being of 'low risk of bias' in the randomisation process, given the random assignment of clusters, and the similarity of cluster sizes and/or balance of household characteristics at pre-test; these included Barrera-Osorio et al. (2014), and Galiani and McEwan (2013) for the replications by Galiani and McEwan (2013) and Galiani et al. (2017).

In Chaplin et al. (2017), the difference in means and statistical tests did not suggest more frequent differences than would be expected by chance alone (9 out of 191 covariates at 5 percent significance). However, there were large differences in baseline variables relating to the outcomes (access to and spending on electricity, use of technologies requiring electricity (e.g., water pump, satellite television), suggesting 'some concerns' which were likely reflected in the small sample size for treatment clusters (27 communities) compared to controls (151 communities). In McKenzie et al. (2010), which compared fewer baseline characteristics, there was a difference in the baseline mean outcome, although that difference was not statistically significant. Nevertheless, it is notable that even small differences may appear significant in relatively large samples, or large differences appear non-significant in small samples. For example, the difference in baseline outcome amounted to 6 percent of the control mean in McKenzie et al. (2010).

In the case of the PROGRESA CCT replications (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006), Behrman and Todd (1999) presented balance tables for several hundred baseline covariates at household level, which suggested statistical differences between treatment and control may have

arisen owing to chance.[118] In contrast, they did not find statistically significant differences in covariates measured at the locality level (where the total sample of treatment and control communities was 505), which suggests that the cluster randomisation led to balanced groups on average. However, no information was available on the randomisation process for PROGRESA – how it was implemented, e.g., with respect to a random number table, and by whom, whether done centrally by researchers – to assess the risk of subversion of randomisation, hence 'some concerns' were noted.

For the RPS CCT programme in Nicaragua, eligible clusters were randomised at a public event, but there appeared differences in group characteristics (the extreme poverty level was higher in controls), as reported in Maluccio and Flores (2005, for the replication by Handa and Maluccio, 2010). This may be due to restricted randomisation over a relatively small cluster sample size (42 clusters in total), which is common in RCT practice.

5.3.3.2 Selection bias

This section assesses risk of selection bias into the study due to identification and recruitment of individual participants in relation to timing of randomisation. It appears the case that individuals were nearly always chosen after randomisation was done or communicated (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006; Handa and Maluccio, 2010; McKenzie et al., 2010; Chaplin et al., 2017). In the case of PROGRESA in Mexico, "[t]he selection of households as PROGRESA beneficiaries was accomplished by first identifying the communities to be covered by the program (geographic targeting) and then selecting the beneficiary households within the chosen communities" (Buddelmeyer and Skoufias, 2004, p.6). Individual household selection was done in a two-part process where eligible households were selected if they fulfilled certain poverty criteria based on a household survey, and then the list presented to the community assembly for discussion, which Skoufias et al. (2001) note made very little difference to the final household choice. As discussed above, although there do not appear to be differences in treatment and control at cluster level, there are differences at the household level, which may go beyond that expected by chance. However, as the authors noted, the large sample size for the study (there were 24,000 households and 41,000 children

[118] Randomisation leads to balanced samples in expectation over repeated trials, not in any specific draw (Edoardo Masset, pers. comm.).

aged under 17) suggested the study was powered to detect very small differences with statistical precision. A more appropriate approach would have been to analyse treatment group and control group differences using distance metrics (Bruhn and McKenzie, 2009), but this was not presented in Behrman and Todd (1999). The benchmark was therefore evaluated as having 'some concerns'.

In the case of RPS in Nicaragua, which used geographic targeting to identify treatment clusters, within which participation was voluntary but participation rates exceeded 90 percent due to the size of the transfer (Handa and Maluccio, 2010), households were chosen for data collection after cluster randomisation, using a random sample based on a household census conducted for the evaluation. Non-response was 10 percent in the first round, and similar in treatment and control groups (Maluccio and Flores, 2005). However, there were differences in baseline household characteristics for a few variables, warranting 'some concerns'.

Similar issues concerning imbalance occurred when assessing McKenzie et al. (2010) and Chaplin et al. (2017). McKenzie et al. (2010) noted difficulties in recruiting individuals into the study, the reasons for which were given for treated units (e.g., being located outside of the survey area) and weighted accordingly, but were less clear for controls. They also attempted to avoid bias in the recruitment strategy which was done by telephoning unsuccessful lottery participants from the same villages as successful participants by including "in the sample households from the Outer Islands of Vava'u and 'Eua" (p.919) that were less likely to have telephones. However, it is not clear how successful the strategy was at obtaining a representative sample of controls, while the reasons for missingness appeared different across treatment and control. In the case of Chaplin (2017), where it appears that sampling of households was done after cluster randomisation, the authors did make efforts to track whether there was migration from controls to treated communities before the household baseline was conducted. However, owing to the differences in baseline characteristics noted above, the analysis suggested 'some concerns'.

In the case of Barrera-Osorio and Filmer (2016), which presented the benchmark RCT used in Barrera-Osorio et al. (2014), recruitment of students, who completed application forms for means-tested scholarships in

treatment and control groups, was done before school-level stakeholders were aware of the school's randomised assignment. In the benchmark study in Galiani and McEwan (2013) and Galiani et al. (2017), all households living in treatment localities were eligible to receive benefits of the programme. In addition, the outcomes data were taken from an unrelated census, conducted 8 months after programme implementation had begun. So, while there could be threats to validity relating to deviation from intended intervention (e.g., due to migration), selection into the study is unlikely to be correlated with treatment status. There was therefore 'low risk of bias' in recruitment of participants for these benchmarks.

5.3.3.3 Attrition

Benchmarks were assessed as having 'low risk of bias' where attrition was at a similar level across treatment and control and where missingness of observations was not differentially correlated with covariates. Studies were of 'some concern' where information was not available. Chaplin et al. (2017) reported data collection in all target communities and 20 percent overall household attrition between baseline and follow-up, evenly split between treatment (19.9%) and control (16.9%), suggesting 'low risk'. The benchmark underlying Galiani and McEwan (2013) and Galiani et al. (2017) was assessed as being of 'low risk of bias' as the analysis was based on census data. McKenzie et al. (2010) performed purposeful sampling of the control group during the follow-up survey because of concerns that the method of follow-up (using a telephone directory) may have led to bias in selection into the study (for those that did not have telephones). They elected to include a sample of participants from the outer islands of Tonga deliberately, in order to correct for the possible bias introduced. However, we remained unclear as to the effect that this purposeful sampling may have had on the composition of the control group and their outcome data during the follow-up. Robustness checks and further details are not available, and therefore the study was rated as having 'some concerns'.

No information was available about differential attrition from the benchmark study for PROGRESA in published reports available. Rubalcava et al. (2009) noted that "one-third of households left the sample during the study period" and "no attempt was made to follow movers" (p.515). No information was reported on differential attrition across groups. PROGRESA was awarded as having 'high risk of bias' due to high overall attrition and lack of information

about differential attrition. In the case of RPS in Nicaragua, there was 5 percent attrition between baseline and follow-up, which was approximately equal in both groups (Maluccio and Flores, 2004). Analysis suggested that attrition may have been correlated with treatment status, but the differences were small, warranting 'some concerns'. In Barrera-Osorio et al. (2014), overall attrition was 23 percent, comprising 20 percent of treated students and 28 percent of controls.

Barrera-Osorio and Filmer (2016) presented significance tests of differences in characteristics between attritors and non-attritors in treatment and control, which they argue are consistent with 'pure chance'. However, due to the differential attrition between groups, the category was classified as having 'some concerns'. Galiani and McEwan (2013) and Galiani et al. (2017) analysed census data for two outcomes – school enrolment, which was available for all households, and child labour, available for 82 percent of households. Although attrition was large for child labour, the data were collected from the census which would not have been linked to the CCT programme by participants or enumerators. Therefore, 'low of risk of bias' was given for attrition.

5.3.3.4 Departures from intended interventions (performance and motivation bias)

Deviations from the intended interventions across the cluster-randomised studies is relevant for within-study comparisons using dependent design, when it affects the control group in the benchmark trial. Issues relating to intervention delays that would typically be of concern if the purpose of the analysis was to estimate treatment effectiveness, are not relevant. For example, referring to the experiment used in Handa and Maluccio (2010), Maluccio and Flores (2004, p.14) stated "it was not possible to design and implement all the components according to the original timelines. In particular, the health-care component was not initiated until June 2001… There were also delays in the payment of transfers to households due to a governmental audit that effectively froze RPS funds." Similarly, Buddelmeyer and Skoufias (2004, p.7) found "in the treatment localities 27% of the total eligible population had not received any benefits by March 2000." However, within-study comparisons based on dependent design estimate the same level of impact, regardless of whether that reflects a poorly implemented intervention. This is particularly relevant for Diaz and Handa

(2006), Handa and Maluccio (2010), Galiani and McEwan (2013), Galiani et al. (2017), Chaplin et al. (2017), and two of the comparisons in Buddelmeyer and Skoufias (2004), where the distance estimate is calculated solely from the comparison of means between randomised control and NRS comparison group. Hence, in these cases, the risk-of-bias rating was amended (upgraded) to capture the expectation that problems in implementation of the intervention would not cause bias between randomised and NRS estimators.

Nevertheless, several cluster-RCTs were considered to have biases in this domain due to potential contamination, spillover effects or performance bias. For PROGRESA (Buddelmeyer and Skoufias, 2004; Diaz and Handa, 2006), Behrman and Todd (1999) explained that individuals may migrate between control and treatment clusters in order to receive the benefits of the intervention and that the incidence of such issues should be tracked. This source of bias may have existed because participating households were not fixed at the start of the study. One of the treatment effect estimates made in Buddelmeyer and Skoufias (2004) suggests potential issues with comparability of the control. In addition, controls within clusters may be affected, where they change their behaviour in response to 'peer effects' from observing treatment participants (spillovers), or possibly with the expectation of becoming eligible for the benefits (John Henry effects) may also have occurred (as also assessed for RPS by Maluccio and Flores, 2005). Buddelmeyer and Skoufias (2003) tested for this by comparing groups where spillovers were unlikely due to geographical separation – i.e., ineligible households in treatment communities compared with eligible but untreated control households (group C versus group B) and ineligible control households (group C versus group D) – and do not find significant differences with the estimates that may have been compromised by spillovers.

However, in general the studies did not indicate the extent that deviations from intended interventions may have occurred. An exception was Maluccio and Flores (2005), which examined the presence of substitution effects in control groups (differential contamination by other interventions) for the RPS CCT programme, finding that there may have been reduced access in control communities for school supplies, but not other interventions. They also reported that a small number of controls who received treatment were dropped from analysis to avoid bias in the estimate. Galiani et al. (2017)

highlighted contamination of controls as an unlikely issue in the benchmark experiment, since the value of the cash transfer was small relative to average income (and there were also severe delays in distribution of the cash transfers beyond the follow-up data collection period). Therefore, the transfers were unlikely to provide incentives or liquidity for poor people to move to treated localities obtain them, in the benchmark study, meriting 'low risk of bias'. Similarly, the scholarship benchmark experiment used by Barrera-Osorio et al. (2014) was assigned 'low risk of bias' on deviation from intended interventions. The analysis used ITT and there were no opportunities for controls to cross over to treatment, since "[i]f a student had dropped out and could not collect the scholarship, the funds could not be reassigned to another student but would be returned to a central fund for use in a subsequent distribution round" (p.473).

Finally, in the case of the natural experiment of the effects of migration on income (McKenzie et al., 2010), there was considerable non-compliance due to no-shows in the treatment group (i.e., a large proportion of participants randomised into the treatment group did not emigrate by the time of the follow-up survey). Two types of experimental estimates were provided by the authors to accommodate deviations from intended interventions. These were ITT, which estimates the effect of assignment, and CACE using instrumental variables, measuring the effect of starting and adhering to treatment, correcting for non-random deviations from the intended intervention.[119] Because the instrument was randomisation, and the correlation with treatment status migration was high (F-statistic=60), this domain was assessed as being of 'low risk of bias'.[120]

[119] The CACE estimate (where the randomised outcome of the random ballot is an instrument for the variable of interest – the migration decision) was the one that was incorporated in subsequent analysis and hence is presented in this analysis.
[120] This is therefore an override to the decision tree used in the Cochrane tool (Higgins et al., 2016) which indicates that even appropriate analysis using IV to correct for non-compliance cannot score more highly than having 'some concerns'. McKenzie et al. (2010) noted: "[v]alidity of the exclusion restrictions then requires: (i) that success in the ballot is uncorrelated with individual attributes which might also affect income, which is provided by the randomization of the ballot draws; and (ii) that the ballot outcome does not directly affect incomes, conditional on migration status. One could conceive of stories such as that winning the ballot and not being able to migrate causes frustration and leads individuals to work less, or conversely, that winning the ballot acts as a spur to work harder in order to afford the costs of trying to find a job in New Zealand. However, we did not encounter any evidence of such changes in behaviour in our field work, lending support to this identification assumption." (p.923)

Table 5.7 Risk-of-bias assessment for within study comparisons

Within study comparison	*Buddelmeyer and Skoufias (2004)**	*Diaz and Handa (2006)**	*Handa and Maluccio (2010)***	*McKenzie et al. (2010)****	*Barrera-Osorio et al. (2014)*****	*Galiani and McEwan (2013); Galiani et al. (2017)******	*Chaplin et al. (2017)*
Confounding bias due to randomisation process	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*	*Some concerns*
Selection bias in recruitment	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*	*Some concerns*
Attrition bias due to missing outcome data	*High risk*	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*
Departures from intended intervention^	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*	*Low risk*	*Low risk*	*Some concerns*
Bias in measurement of the outcome^	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*	*Low risk*
Selective analysis and reporting^	*Low risk*	*Low risk*	*Low risk*	*Low risk*	*Low risk*	*Low risk*	*Low risk*
Bias in NRS estimate	*Low risk*	*Some concerns*	*Some concerns*	*Low risk*	*Low risk*	*Some concerns*	*Low risk*
Overall bias in within-study comparison	*High risk*	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*	*Some concerns*

Notes: * assessment draws on Behrman and Todd (1999), Skoufias et al. (2001), Angelucci and de Giorgi (2006) and Rubalcava et al. (2009); ** assessment draws on Maluccio and Flores (2004, 2005); *** assessment is of the instrumental variables estimate for the randomised sample (complier average causal effect); **** assessment draws on Barrera-Osorio and Filmer (2016); ***** assessment draws on Glewwe and Olinto (2004); ^ assessment takes into account relevance of the domain for relative bias regarding within-study comparison.
Source: author using Higgins et al. (2016), Eldridge et al. (2016) and the critical appraisal tool developed in Chapter 4 and presented in Appendix A.

5.3.3.5 Bias in measurement of outcome

While assessment of confounding and differential selection bias (into and out of study) are important in determining absolute bias in the benchmark estimate itself, as well as bias in relation to the NRS estimate, it is not immediately clear whether risk of bias in the method of collecting outcomes data is important in determining the relative bias between them. For example, if outcomes data were collected using identical methods (whether observed or reported) in an open benchmark control and NRS comparison study, these potential biases might be expected to 'cancel out' in the calculation of the distance estimate. Whether this is the case would depend on the motivations of participant or outcome assessors in unblinded studies, which may vary between trials where data are clearly linked to an intervention (due to informed consent), and studies where data are not. Hence, in the case of McKenzie et al. (2010), an individually randomised lottery, where benchmark and NRS outcomes data were collected using the same tools by the same enumerators, it is possible that migrants were incentivised to over-report income (e.g., due to the 'false success' narratives that are known to exist) (Waddington and Sabates-Wheeler, 2002), which could have upwardly biased the benchmark estimate.[121]

Across all but three benchmarks, outcome measurements were considered to have 'some concerns'. This was largely due to the issue of lack of blinding of assessors in trials, where participants and outcome assessors may have had incentives affecting how they report outcomes (e.g., relating to social desirability). It is also unknown (there was insufficient evidence) to confidently state whether outcomes were likely to be influenced by knowledge of intervention received, since outcomes data were usually collected from household surveys through self-report, rather than more rigorous methods such as formal tests.[122] However, these are also cluster-RCTs where informed consent for the outcomes survey does need not refer to a specific intervention. Unfortunately, no information was reported about the process of consent in any of the studies, so it was unclear whether consent

[121] McKenzie et al. (2010) also collected pre-test income using one-year recall, which might be expected to be less reliable for international migrants than non-migrants. However, the treatment effect estimates calculated for benchmark and NRS in this study do not use the pre-test outcome.

[122] In some instances, outcomes were collected at community level (e.g., household electricity grid connections in Chaplin et al., 2017) but these were not used in estimation of within-study comparisons.

for the benchmark studies informed participants that the purpose of data collection was to evaluate the intervention of interest.

Outcomes were observed for Barrera-Osorio et al. (2014), where enumerators determined grade completion and administered mathematics tests. The data were collected using the same survey instrument at the same time in benchmark and NRS comparison. Therefore, even though outcome assessors were not blinded, any effect that enumerator incentives may have had was likely to be equivalent in benchmark control and NRS comparison. In the benchmark study in Galiani and McEwan (2013) and Galiani et al. (2017), there was effectively blinding of outcome assessment, since outcomes data used the national census which had been collected shortly after implementation of the cash transfer programme. Participants and outcome assessors would therefore not have been able to associate the data collection with the programme or household treatment status. Both studies were therefore rated as having 'low risk of bias' in outcomes measurement.

In the case of Chaplin et al. (2017), outcomes data were collected through self-report in benchmark and NRS using the same survey instruments at the time of year by the authors. Furthermore, the NRS comparison group was selected from the comparison group of a concurrent non-randomised evaluation of electrification being done by the authors for the same project at the same time as the RCT. Therefore, since both benchmark control and NRS comparison data were from communities taking part in evaluations, any effect that knowledge of treatment status by participants or enumerators may have had on responses may be expected to 'cancel out'. Hence, this study was also assigned 'low risk of bias'.

5.3.3.6 Selective analysis and reporting

The purpose of the within-study comparisons was usually to test for differences across multiple outcomes and specifications of benchmark and NRS comparison. Selective reporting was assessed as being of 'low risk bias' across all benchmark studies, due to the large number of effects usually reported for different outcomes and samples. For example, all studies reported results of RCTs across multiple outcome domains, which were subsequently used in comparison with non-randomised replications. Some studies also reported findings for particular sub-groups, such as boys and girls in Buddelmeyer and Skoufias (2004), which was judged as common

practice in the evaluation of school programmes, and non-selectively reported since all findings were reported by sex for all specifications. However, there is potentially a problem with multiple hypothesis, suggesting that statistical significance thresholds should be more conservative when comparing differences between RCT and NRS.

5.3.3.7 Bias in the within-study comparison estimate

The final source of bias relates to confounding of the relative estimates of benchmark and NRS due to differences in measurement and differences in target population (sampling bias). This section discusses these sources of bias, as well as threats to validity due to implementation of the NRS (Table 5.8 provides a detailed summary, which is presented as an overall rating in Table 5.7). Regarding measurement, McKenzie et al. (2010) reported NRS findings for two surveys, one done by the authors identical to that done for the randomised benchmark, comprising a relatively small sample of 60 non-applicant households living in the same village as lottery applicants. The second survey drew on nationally representative survey containing 3,000 households in the relevant target population. The findings reported below are therefore taken from the author survey to ensure identical survey instruments.[123]However, Diaz and Handa (2006) reported differences in sampling frame and season of data collection between benchmark and NRS for all outcomes, as well as specific differences in detail of questions and recall period for expenditure data, stating that "differences in expenditure outcomes may be entirely due to questionnaire design rather than evaluation technique" (p.327). These differences were noted and explored in meta-analysis below.

Handa and Maluccio (2010) used Living Standards Measurement Survey (LSMS) data, which were collected at the start of the rainy season in April to July 2001, to generate the NRS comparison for RCT data collected in October 2001, at the end of the rainy season. Given likely seasonal variation in the outcomes measured (food expenditure, preventive health care behaviour, child health), it was useful both surveys were done in the same season, although it is possible the RCT data were collected at the time when

[123] The findings from the nationally representative survey data were reported for OLS and PSM specifications in McKenzie et al. (2010). These yielded distance metrics larger than for the survey data collected by the authors (reported in Table 5.9) and Table 5.10. The mean distance for OLS specifications is 0.104 (95%CI=0.013, 0.194); for PSM it is 0.096 (95%CI=-0.021, 0.214).

infectious diseases (e.g., diarrhoea and ARIs) were more prevalent, which would tend to cause the mean in the RCT control to exceed the NRS comparison.[124] Furthermore, for one of the 12 outcomes collected, use of preventive health check-up for children aged 0-36 months, there were slight differences in the reference period being recalled and specific type of check-up. However, the authors made refinements to the LSMS sample used in the NRS to foster comparability with the RCT sample. Firstly, they excluded localities where the programme was operating from the NRS sample, to avoid possible contamination from treated households (since the programme began the previous year). From this sample, they calculated three NRS treatment estimates: the full sample estimate; a sub-sample estimate including only those localities that would have been eligible for treatment using the marginality index that determined eligibility for treatment; and a second sub-sample limiting eligible localities to the same geographical zone as treated households. Differences in findings for these sub-samples were explored in the meta-analysis below.

A second question is whether there are differences in the NRS treatment estimand (e.g., ATET or LATE) with the benchmark (estimating ATE) that would lead to differences in treatment quantity over and above any bias or sampling error. In nearly all cases, the authors ensured NRS target populations were as similar as possible to RCTs, or the bias estimates were able to incorporate the differences. For example, in all RDD within-study comparisons, the RCT results were estimated at the same bandwidth around the treatment threshold. In the matched NRS comparisons, the bias was calculated with reference to the RCT control group only (Diaz and Handa, 2006; Handa and Maluccio, 2010), hence adjustments based on non-compliance were not necessary.[125] However, in McKenzie (2010) the bias estimates relied on the treatment mean. There was substantial non-compliance with the migrant lottery (mainly due to delays in migration). Therefore, the complier average causal effect (CACE) estimate using

[124] The rainy season in Nicaragua is from May to October, with the wettest months being September and October.
[125] For example, in Handa and Maluccio (2010), the benchmark effect estimand was the intention-to-treat. The intervention participation rate was 90 percent, however, suggesting that NRS estimates of treatment effect, using ATET, would need to be rescaled by dividing ATET by (1-0.9) = 0.1, in order to equivalise the denominator and ensure comparability, if the analysis were comparing the treatment effect estimates.

instrumental variables was taken for the benchmark estimate, rather than the ITT estimate.

Table 5.8 Bias in NRS-RCT comparisons

Within study comparison (outcome)	*Risk rating*	*Cause of confounding in NRS-RCT bias estimate*
Buddelmeyer and Skoufias (2004)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold
Diaz and Handa (2006) (education)	Some concerns	Difference in season and sampling frame between NRS and benchmark surveys
Diaz and Handa (2006) (expenditure and child labour)	High risk	Measurement of expenditure and child labour differ between NRS and RCT surveys
Handa and Maluccio (2010) (expenditure, child feeding practices, immunisation)	Low risk	NRS and benchmark use same survey questions and target populations, during same season
Handa and Maluccio (2010) (child illness in previous month)	Some concerns	NRS and benchmark surveys conducted at opposite ends of the rainy season
Handa and Maluccio (2010) (preventive health)	High risk	NRS and benchmark questions are different for preventive health.
McKenzie et al. (2010)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold
Barrera-Osorio et al. (2014)	Low risk	NRS and RCT use same survey and bandwidth around eligibility threshold
Galiani and McEwan (2013)	Some concerns	NRS and RCT use same survey and bandwidth around eligibility threshold; some concerns about the method used to identify the NRS comparison group.
Galiani et al. (2017)	Some concerns	NRS and RCT use same survey and bandwidth around eligibility threshold; some concerns about the comparability of the NRS population.
Chaplin et al. (2017)	Low risk	NRS and RCT use same survey conducted during same time of year

Regarding the implementation of the NRS, the studies reported sensitivity analysis using different estimators. For example, the studies of matching assessed the differences with nearest-neighbour, caliper, kernel and local-linear algorithms (e.g., Diaz and Handa, 2006), the inclusion of baseline outcome (McKenzie et al., 2010; Chaplin et al., 2017), use of 'rich covariates' and geographically proximate observations (Handa and Maluccio, 2010; Chaplin et al., 2017). The study of instrumental variables examined

sensitivity to alternative instruments (McKenzie et al., 2010). Studies of regression discontinuity compared different bandwidth estimates (Buddelmeyer and Skoufias, 2004; Barrera-Osorio et al., 2014). Whether these differences are correlated with bias is an empirical question that was explored in the meta-analysis below.

However, it was also important to consider the quality of implementation in the NRS. For example, matching should be done using covariates that are likely to be correlated with treatment and outcome, preferably using higher-order polynomials and interactions with the treatment variable (Handa and Maluccio, 2010), but importantly the covariates must not be affected by the treatment. Handa and Maluccio (2010) used locality variables measured five years prior to treatment (which could not have been affected by treatment), and household variables measured one year after treatment commenced, some of which were fixed (e.g., age and parental education) but others may have been affected (e.g., working patterns). By contrasting the findings with NRS matches made using a survey from the previous year, they interpreted the findings as presenting evidence of bias in some of the household level matching variables.

Matches should also not be so geographically proximate as to lead to possible bias in the treatment effect due to contamination or spillovers. Two points may be noted here. The first is that in nearly all cases, bias is estimated exclusive of treatment observations by comparing benchmark control and NRS comparison means, so there is no risk of contamination. In the case of McKenzie et al. (2010) where bias is calculated using the treatment mean as well, and NRS comparisons are taken from the same communities where treated observations used to live, there is also little risk of contamination owing to the nature of the intervention (international migration). It is also worth noting that 'geographical proximity' is fairly loosely defined, as coming from the same central part of the country in Handa and Maluccio (2010). In Tanzania, Chaplin et al. (2017, p.G.7) stated they "initially concluded that 30 km would be a reasonable radius based on the following criteria: that 30 kilometers is an upper bound for the distance most adults would reasonably

walk in a day and used it as one measure of how much two communities would be subject to similar influences."[126]

In the scholarship RDD (Barrera-Osorio et al., 2014), assignment was based on one of two indexes – a merit threshold based on a student test score, and a poverty threshold based on students' reported household and family socioeconomic factors. The tests were scored centrally by an independent firm employed specifically to reduce manipulation of eligibility. The authors noted that the official list of scholarship recipients provided by the government was identical to the list provided by the firm. Furthermore, "spot checks at a number of schools yielded no cases of the manipulation of the selection process" (Barrera-Osorio and Filmer, 2014, p. 486).

In Galiani and McEwan (2013), precise HAZ-score programme eligibility data were only available for the benchmark localities. However, a report on the height census conducted four years previously gave the proportion of children with severe and moderate stunting (HAZ-scores below -3 and -2, respectively) for all localities nationally. Eligibility for the RDD comparison localities were then predicted by a regression of the mean HAZ-score from the censored data on the stunting proportions from the previous height census. The authors found a high correlation between predicted HAZ-score and actual HAZ-score for treatment communities (r=0.96), although it should be borne in mind that eligibility for the RDD comparison is therefore estimated and 'fuzzy'. In Galiani et al. (2017), there were also concerns in the design of the NRS replication due to the "persistent imbalance in one covariate (Lenca) that is plausibly correlated with unobserved determinants of child outcomes" (p.207) between treated and control municipalities. As the authors argued, it was therefore not possible to assume continuity in potential outcomes at municipal borders, suggesting some threats to internal validity of the replication. Therefore, despite the benchmark in Galiani and McEwan (2013) and Galiani et al. (2017) being assessed as of 'low risk of bias', concerns about implementation of the NRS suggested 'some concerns' about confounding of the difference estimator.

[126] However, Chaplin et al. (2017) also discussed the potential limitations of local matching on reducing the availability, and therefore quality, of potential matches, and settled on a radius of 40 kilometres.

5.3.4 *Quantitative estimates of bias*

Data were collected on treatment effects for the benchmark study, as well as each corresponding non-randomised replication from 545 specifications. These data included outcome means in treatment and control/comparison (or treatment effect estimates from an analysis), outcome variances, sample sizes and significance test values (e.g., t-statistics, confidence intervals, p-values). The estimate of effect which most closely corresponded with the population for the non-randomised arm was taken from the RCT – the bandwidth around the treatment threshold in the RDDs (Buddelmeyer and Skoufias, 2004; Galiani and McEwan, 2013; Barrera-Osorio et al., 2014; Galiani et al., 2017), and the instrumental variables analysis of the randomised natural experiment in McKenzie et al. (2010).

Five distance metrics were used to compare the difference between NRS and benchmark means, interpreted as the magnitude of bias in the NRS estimator: the standardised difference and the percentage difference (Steiner and Wong, 2016); the absolute difference as a percentage of the control mean (Glazerman et al., 2003); the percentage reduction in bias (Chaplin et al., 2017); and the mean squared error (e.g., Greenland, 2000). As above, *D* is defined as the primary distance metric measuring the difference between the non-experimental and experimental means, interpreted as the size of the bias, calculated as:

$$D = \hat{\tau}_{NRS} - \hat{\tau}_{RCT} = (\bar{Y}^{c}_{NRS} - \bar{Y}^{t}_{RCT}) - (\bar{Y}^{c}_{RCT} - \bar{Y}^{t}_{RCT}) = \bar{Y}^{c}_{NRS} - \bar{Y}^{c}_{RCT} \qquad (5.6)$$

where $\bar{Y}^{c}_{NRS}$ and $\bar{Y}^{c}_{RCT}$ are the mean outcomes of the non-randomised comparison and randomised control groups, and $\bar{Y}^{t}_{RCT}$ is the mean outcome of the randomised treatment group. Taking the absolute difference in *D* ensures consistency across studies' reported effects, since a large number of values of *D* were collected from each study. This ensured that a measure of the overall deviation of randomised and non-randomised estimators was estimated, and not a measure that, on average 'cancelled out' positive and negative deviations, potentially obscuring differences of interest.[127] Following Steiner and Wong (2016), the standardised absolute difference

[127] In practice, standardised difference from the simple subtraction of RCT from NRS estimate was frequently either side of zero, which did tend to 'cancel out' across specifications, as shown in the results.

$|D_s|$ between treatment effects in experimental and non-randomised replication samples was calculated:

$$|D_S| = \frac{|\bar{Y}^c_{NRS} - \bar{Y}^c_{RCT}|}{S_{RCT}} \qquad (5.7)$$

where S_{RCT} is the sample standard deviation of the outcome in the benchmark study. Where the standard deviation was not reported, it was calculated from reported data using formulae in Appendix C.[128] If the benchmark study did not report the standard deviation of the outcome, but the standard error *se(b)* of the test statistic for effect size estimate *b* was available, as in the case of McKenzie et al. (2010), the standard deviation was calculated using (Borenstein et al., 2009):

$$S \quad = se(b)\sqrt{\frac{n_t\, n_c}{n_t + n_c}} \qquad (5.8)$$

Where group sample sizes can be assumed equal this simplifies to:

$$S \quad = se(b)\sqrt{\frac{N}{4}} \quad if\ n_t = n_c = \frac{N}{2} \qquad (5.9)$$

which was used to calculate the outcome standard deviation in Galiani and McEwan (2013), Barrera-Osorio et al. (2014) and Galiani et al. (2017).

The standard error of D_s is given by:

$$se(D_S) = \sqrt{se^2_{NRS} + se^2_{RCT}} \qquad (5.10)$$

where se_{NRS} and se_{RCT} are the standard errors of the non-randomised and randomised mean outcomes, respectively, which can be assumed independent. The test statistic is given by:

[128] For example, in the case of Handa and Maluccio (2010) the outcome standard deviations for proportion effect sizes were not reported, but information on the treatment effect, control mean and sample sizes were. The standard deviation of the outcome was calculated from this information using equations A.4-A.6 in Appendix 3.

$$|t| = \frac{|D_s|}{se(D_s)} \qquad (5.11)$$

and 95 percent confidence interval:

$$|D_S| \pm 1.96 * se(D_s) \qquad (5.12)$$

These calculations were made for all studies apart from Chaplin et al. (2017) who reported average differences across outcomes standardised by the randomised control mean. In addition, several studies used boot-strap methods to generate the variance for matched comparisons (Diaz and Handa, 2006; Handa and Maluccio, 2010; McKenzie et al., 2010), which were used to calculate the confidence intervals.[129]

For comparison purposes, the standardised numerical difference was also calculated, since individual studies (and the existing review in development economics by Hansen et al., 2013) seem to have used it in generalising findings:

$$D_S = \frac{\bar{Y}^c_{NRS} - \bar{Y}^c_{RCT}}{S_{RCT}} \qquad (5.13)$$

The standard error of D_s uses equation (5.10). However, the limitation of using the standardised numerical or absolute distance is that is it not easily interpretable. Therefore, following Steiner and Wong (2016), bias was also calculated as a percentage of the RCT treatment effect estimate:[130]

$$|D_T| = \frac{\hat{\tau}_{NRS} - \hat{\tau}_{RCT}}{|\hat{\tau}_{RCT}|} x\ 100 = \frac{\bar{Y}^c_{NRS} - \bar{Y}^c_{RCT}}{|\bar{Y}^t_{RCT} - \bar{Y}^c_{RCT}|}\ x\ 100 \qquad (5.14)$$

However, the limitation of this approach is that the percentage difference can become very large where the benchmark estimate is close to zero (Steiner and Wong, 2016), and the estimator not identified when the benchmark estimate

[129] In the case of Handa and Maluccio (2010), there appears to be misreporting in tables 1-3 of that study reporting bias estimates for nearest neighbour and kernel matching. Therefore, in this case, bias was re-estimated using information reported on RCT and NRS treatment impacts, and the standard error of the bias was re-estimated using the t-statistic of the reported bias estimate.

[130] Briscoe et al. (1985) proposed a similar normalised bias estimator for infectious disease morbidity in WASH studies: $Bias = \frac{\widehat{OR} - OR^*}{OR^* - 1}$, where $\widehat{OR}$ is the observed odds ratio and OR^* is the 'true' odds ratio measured without bias.

is equal to zero. With the aim of providing an interpretable benchmark, bias was also calculated as the percentage of the control mean (Glazerman et al., 2003), with the caveat that a control mean that is close to zero may also generate a big percentage:

$$|D_C| = \frac{|\bar{Y}^c_{NRS} - \bar{Y}^c_{RCT}|}{\bar{Y}^c_{RCT}} \; x \; 100 \qquad (5.15)$$

An estimator of bias used in Chaplin et al. (2017), was slightly modified from the 'percentage of remaining bias' estimator defined by Steiner and Wong (2016):

$$|D_R| = \left(1 - \frac{\bar{Y}^c_{NRS} - \bar{Y}^c_{RCT}}{|\bar{Y}^c_{PF} - \bar{Y}^c_{RCT}|}\right) x \; 100 \qquad (5.16)$$

which estimates the percentage of bias removed by the NRS, where $\bar{Y}^c_{PF}$ is the *prima facie* comparison mean from the unadjusted non-randomised model. The data were also available to calculate this estimator in Diaz and Handa (2005). It was possible to calculate a 'percentage of remaining bias' estimator for other studies (Buddelmeyer and Skoufias, 2004; McKenzie et al., 20100; Handa and Maluccio, 2010; Galiani, 2013, 2017), by using the pre-test mean as the *prima facie* estimator.[131] Where only the treatment group pre-test post-test difference was available, $\bar{Y}^t_{RCT_1} - \bar{Y}^t_{RCT_0}$, the benchmark treatment effect $\bar{Y}^t_{RCT_1} - \bar{Y}^c_{RCT_1}$ was subtracted from it, to obtain the relevant quantity for the denominator in equation (5.14):

$$|\bar{Y}^c_{PF} - \bar{Y}^c_{RCT}| = |(\bar{Y}^t_{RCT_1} - \bar{Y}^t_{RCT_0}) - (\bar{Y}^t_{RCT_1} - \bar{Y}^c_{RCT_1})| = |\bar{Y}^c_{RCT_1} - \bar{Y}^t_{RCT_0}| \qquad (5.17)$$

Finally, in order to facilitate comparisons across NRS estimates, the expected mean squared error was calculated for each distance estimate:

$$MSE_i = bias_i^2 + s_i^2 = D_i^2 + s_i^2 \qquad (5.18)$$

The initial results used averaging over the large number of values of *D* collected in each study. Mean standardised bias estimates reported for each

[131] For Handa and Maluccio (2010), data were available for six variables (total, adjusted and food expenditure, up-to-date immunisation and health check-ups) in Maluccio and Flores (2005). For Galiani and McEwan (2013) and Galiani et al. (2017), data were available on education enrolment in Glewwe and Olinto (2004).

included study are reported for regression studies (Table 5.9), matching studies (Table 5.10) and discontinuity designs (Table 5.11). These tables use simple averages from the 545 individual standardised numerical differences and standardised absolute differences of mean bias and their standard errors. The findings also accounted for differences in implementation of the NRS comparison, as well as issues that threatened the comparability of the NRS and benchmark (e.g., outcome estimate, treatment estimand).

Table 5.9 Mean standardised bias estimates in regression studies

Study	*NRS type*	*Standardised numerical difference*	*Standardised absolute difference*	*Num. bias estimates*
Diaz and Handa (2006)	OLS	0.257	0.262	6
McKenzie et al. (2010)	OLS	0.195	0.195	4
	IV	0.206	0.206	3
	IV (valid instrument)	0.007	0.007	1
	DD	0.137	0.137	2

Two within-study comparisons reported distance using regression-based estimators (Diaz and Handa, 2006; McKenzie et al., 2010) (Table 5.9). The OLS specifications may perhaps be one benchmark against which other estimators may be compared. As expected, OLS distance estimators tended to be larger than those using other methods, including double differences, valid instrumental variables and matching. McKenzie et al. (2010) also reported the single difference estimator, taken from the difference between pre-test and post-test outcome, equal to 0.156. This was found to be a less accurate predictor of the counterfactual outcome than matching including baseline outcome, double differences and instrumental variables estimation using effective instruments, but more accurate than OLS and statistical matching which excluded the baseline outcome (Table 5.10).

Table 5.10 Mean standardised bias estimates in matching studies

Study	*NRS type*	*Standardised numerical difference*	*Standardised absolute difference*	*Num. bias estimates*
Chaplin et al. (2017)*	Matching	-	0.103	7
	Matching (local comparison, L)	-	0.091	1
	Matching (pre-test outcome, P)	-	0.085	1
	Matching (rich controls, C)	-	0.033	1
	Matching (LPC)	-	0.024	1
Diaz and Handa (2006)	Matching	0.168	0.242	24
	Matching (parsimonious controls)	0.390	0.394	24
	Matching (education enrolment)	-0.057	0.066	8
Handa and Maluccio (2010)	Matching	-0.023	0.319	132
	Matching (local comparison)	-0.028	0.235	45
	Matching (reported child illness)	-0.044	0.145	8
McKenzie et al. (2010)	Matching	0.151	0.151	15
	Matching (including pre-test outcome)	0.143	0.143	6

Notes: * study presents mean estimates from distance estimates conducted for 59 outcome variables; - estimator not calculable.

McKenzie et al. (2010) examined the use of two-stage least squares (2SLS) instrumental variables estimation, studying the effects of immigration on income using NRS data. One instrument was the migrant's network (indicated by number of relatives in the country of immigration). This was shown to be correlated with migration (albeit by an F-statistic=6, below the satisfactory threshold of F=10), but produced a treatment effect distance

metric that exceeded single differences (pre-test post-test) and OLS, supporting the theoretical prediction that inappropriate instruments produce 2SLS findings that are less consistent than OLS (Wooldridge, 2009). The authors argued it was unlikely to satisfy the exclusion restriction since it was very likely correlated with income after immigration, despite being commonly used in the field of migration. Another instrument, distance to the application centre, also frequently in instrumental variables, produced the smallest distance metric of any within-study comparison, effectively equal to zero. The instrument was highly correlated with migration (F-statistic=40) and, it was argued, was satisfied the exclusion restriction as it was unlikely to determine income for participants on the main island where "there is only a single labor market… where all villages are within one hour of the capital city" (p.939). However, it also not possible to rule out the possibility that the arguments being made for success of the instrument were based on results. Distance is usually seen as a weak instrument, since programme participants can move to obtain access to services. The other point worth noting is that IV produced mean squared error greater than OLS, whether it was estimated using valid or invalid instruments, owing to the greater imprecision of 2SLS estimation.

Four studies estimated distance using statistical matching (Diaz and Handa, 2006; Handa and Maluccio, 2010; McKenzie et al., 2010; Chaplin et al., 2017). Matching estimators tended to be relatively large on average (between 0.10 and 0.30 in simple specifications) (Table 5.10). The bias coefficients were smaller when using more advanced approaches, including pre-test outcomes, local matches and rich specifications.

It is worth remembering that the pre-test outcome in McKenzie et al. (2010) was measured through one-year recall, which may be liable to bias, hence the pre-test outcome matching estimator does not substantially affect the bias estimate. In addition, McKenzie et al. (2010) implicitly used local matches, by choosing NRS comparisons from geographically proximate households in the same villages as treated households.[132]

[132] Due to the reduced risk of contamination, as the treated households had emigrated already, matches in McKenzie et al. (2010) could be from the same villages, unlike in other matched studies (for an intervention where there is a risk of contamination or spillover effects), where matches would need to be geographically separate.

Two studies showed that more parsimonious matching (reducing the covariates in the matching equation to social and demographic characteristics that would be available in a typical household survey) estimated bigger distances from benchmark than matching using rich control variables in the data available (Diaz and Handa, 2006; Chaplin et al., 2017). Two studies also showed that matching on pre-test outcomes also provided smaller distance metrics (McKenzie et al., 2010; Chaplin et al., 2017). Finally, two studies showed smaller distance estimates when matching on local comparisons (Handa and Maluccio, 2010; Chaplin et al., 2017).

In Diaz and Handa (2006) there was an additional source of confounding in the distance estimate due to differences in survey questionnaire for expenditure and child labour outcomes. When the outcome was restricted to education enrolment which was measured comparably across surveys, the distance estimator was less than 0.1 standard deviations. In Handa and Maluccio (2010), the smallest distance estimate was for reported child illness.

Table 5.10 also clearly demonstrates that the calculation of difference, using numerical or absolute values, can lead to very different mean bias values, where the individual underlying difference estimates are distributed above and below the null effect. Essentially, using absolute mean differences accentuates the difference between RCT and NRS mean, and will always be greater than zero.

Four studies examined discontinuity designs (Buddelmeyer and Skoufias, 2004; Galiani and McEwan, 2013; Barrera-Osorio et al., 2014; Galiani et al., 2017), producing distance metrics that were typically less than 0.1 standard deviations (Table 5.11). These relatively small distance metrics, compared with the other NRS estimators, varied by the bandwidth used (Buddelmeyer and Skoufias, 2004). In Barrera-Osorio et al. (2014), the bias in test scores estimates was substantially smaller than the bias in grade completion, which the authors noted was estimated by enumerators and may therefore have been measured with bias.

Table 5.11 Mean standardised bias estimates in discontinuity designs

Study	*NRS type*	*Standardised numerical difference*	*Standardised absolute difference*	*Num. bias estimates*
Barrera-Osorio et al. (2014)	RDD	-0.062	0.119	20
	RDD (grade completion)	-0.176	0.184	10
	RDD (math test)	0.053	0.054	10
Buddelmeyer and Skoufias (2004)	RDD	-0.017	0.073	214
	RDD – narrow bandwidth	-0.030	0.060	72
	RDD – medium bandwidth	-0.029	0.083	72
	RDD – wide bandwidth	-0.031	0.077	72
Galiani and McEwan (2013)	RDD	0.003	0.008	9
Galiani et al. (2017)	GDD	0.008	0.018	72

However, the estimates presented here were calculated using simple averages and, fundamentally, it remains unclear whether the differences are substantively important. In order to account for differences in precision, pooled means across studies were calculated using fixed effect inverse variance-weighted meta-analysis. The fixed effect model may be justified under the assumption that the estimates are from the same target populations, with the remaining bias being due to sampling error. However, each internal replication study reported multiple bias estimates using different methods of analysis and/or specifications. The weights w for each estimate needed to take into account the different numbers of bias estimates each study contributed, using the following approach:[133]

[133] Following Hedges et al. (2010), a generalised approach is presented in Tanner-Smith and Tipton (2014):

$$w_{ij} = \frac{1}{(s_i^2 + \tau^2)[1 + (m_j^k - 1)\rho]}$$

where the weighting takes into account the between-studies error in a random effects model, τ^2 (equal to zero in the fixed effect case), and the estimated correlation between effects, ρ (equal to 1 where all NRS comparisons draw on the same sample and the benchmark control is the same across all distance estimates).

$$w_{ij} = \frac{1}{s_i^2} \cdot \frac{1}{m_j^k} \qquad (5.19)$$

where s_i^2 is the variance of distance estimate i and m_j^k is the number of distance estimates provided by study k. The pooled weighted average of D was calculated as:

$$D = \frac{\sum_{ij} w_{ij} D_{ij}}{\sum_{ij} w_{ij}} \qquad (5.20)$$

Noting that the weight for a single study is equal to the inverse of the variance for each estimate adjusted for the total number of estimates, following Borenstein et al. (2009), it follows that the variance of the weighted average is the inverse of the sum of the weights across k included studies:

$$s_D^2 = \frac{1}{\sum_{ij} w_{ij}} \qquad (5.21)$$

Table 5.12 compares the distance estimates obtained from different methods of calculating the pooled effect. It can be seen that the simple average of the subtraction of NRS from RCT mean tends to underestimate the distance metric, by generating positive values that on average 'cancel out' (Table 5.12, column 1). The corollary is that, taking the simple average of the absolute difference produces distance estimates that tend to be bigger (Table 5.12, column 2). This explains why the findings from this review are different from those found in the original within-study comparison papers, which implicitly used averaging of the subtraction in discussion of their findings. However, this method may overestimate the typical distance metric.

On the other hand, using the adjusted inverse-variance weighted average, produces distance metrics between these two extremes (Table 5.12, column 3). Even the metrics for matching are below 0.1 in these cases, although this is due to the large number of small distance metrics produced by Chaplin et al., 2017). When the studies are instead weighted by RCT sample size,[134]

[134] Sample size weighting uses the following formula: $w_{ij} = n_i \ / m_j^k$ where n_i is the sample size for difference estimate i and m_j the number of estimates contributed by study k.

rather than inverse of the variance, the matching distance metrics revert to magnitudes presented above (Table 5.12, column 4). The remaining columns use the sample size weights (adjusted for the number of estimates contributed by the study, as above).

The remaining columns attempt to translate the findings into metrics that better indicate the substantive importance of the bias which is represented by these distance estimates. Table 5.12 Column 5 gives the mean squared error, column 6 presents the bias as a percentage of the benchmark treatment effect, and column 7 gives bias as a percentage of the benchmark control mean. For example, RDD estimation produces bias that is on average different from the RCT treatment effect by 7 percent, and 8 percent of the control mean. However, when RDD is compared to ATE estimates, it produces distance estimates that are on average 20 percent different from the RCT estimate. These findings were strengthened by the inclusion of distance estimates from two studies that were excluded from previous analysis (Urquieta et al., 2009; Lamadrid-Figueroa et al., 2013), which compared RDD estimates with RCT ATEs. Regarding statistical significance of the findings, RDDs are also usually of lesser power because they are estimated for a local population around the cut-off.[135] However, mean squared error estimates tend to be relatively small in comparison with other NRS estimators.

[135] For example, Goldberger (1972) originally estimated sampling variances for an early conception of RDD as being 2.75 times larger than an RCT of equivalent sample size.

Table 5.12 Pooled standardised bias estimates

Estimator	(1) Simple subtracted standardised bias*	(2) Simple absolute standardised bias*	(3) Absolute standardised bias**	(4) Absolute standardised bias***	(5) Mean squared error***	(6) Percent difference***	(7) Percent bias***	(8) Percent bias removed***	Number of distance estimates$
OLS	0.232	0.236	0.229	0.290	0.180	340.8	26.3	33.9	10
RDD (LATE)^	-0.015	0.048	0.025	0.012	0.000	7.3	8.0	94.2	173
RDD (ATE)^	-0.057	0.091	0.038	0.029	0.002	N/A	N/A	N/A	71
IV	0.206	0.206	0.104	0.206	0.095	31.8	83.8	-29.6	3
IV (strong instrument)	0.007	0.007	0.007	0.007	0.000	1.1	3.0	95.4	1
IV (weak instrument)	0.305	0.305	0.184	0.305	0.142	47.2	124.2	-92.1	2
DD	0.137	0.137	0.137	0.137	0.019	21.2	55.9	55.9	2
Matching	0.084	0.280	0.059	0.246	0.183	210.3	13.1	58.3	177
Matching on local comparison	-0.001	0.200	0.043	0.088	0.028	-9.0	-8.4	53.1	66
Matching on baseline outcome	0.120	0.120	0.039	0.044	0.004	1.6	4.3	55.9	15
Matching with rich controls	0.025	0.282	0.031	0.232	0.124	178.4	8.4	81.3	116
Parsimonious matching	0.208	0.321	0.087	0.354	0.305	353.8	25.9	44.9	44
Nearest neighbour matching	0.011	0.273	0.040	0.125	0.069	54.9	-2.4	51.5	59
Kernel matching	0.022	0.284	0.139	0.282	0.149	255.3	4.3	34.3	70
Local linear matching	0.179	0.257	0.201	0.285	0.133	267.8	18.3	21.4	6
Radius matching	0.154	0.256	0.215	0.280	0.143	100.8	11.0	236.6	6

Notes: * simple average used to calculate pooled estimate; ** weighted average calculated using the inverse of the variance multiplied by the inverse of the number of estimates in the study; *** weighted average calculated using the benchmark sample size multiplied by the inverse of the number of estimates in the study; ^ indicates RDD estimate compared with either RCT local average treatment effect or RCT average treatment effect (ATE comparisons also incorporated Urquieta et al., 2009 and Lamadrid-Figueroa et al., 2013); $ sample size is for calculations in (1-7), calculations in (8) use a smaller number of studies owing to more limited availability of a *prima facie* estimate.

IV using strong instruments is even more accurate, although only two estimates were available in the studies. Matching tends to produce estimates that differ from the RCT treatment effect by large percentages, on average twice that of the RCT estimate. However, matching would be expected to present a larger treatment estimate where it estimates ATET, which is bigger than ATE under non-adherence. Presenting bias as a percentage of the control mean, the estimates are smaller. However, as noted above, where the control mean is close to zero, or small relative to the treatment estimate, the percentage difference estimator can be large, as was the case in many of the matching estimators presented in Handa and Maluccio (2010). The mean squared errors for matching also tended to be smaller than comparable estimates using OLS (Diaz and Handa, 2006; McKenzie et al., 2010).

5.3.5 *Conclusions*

A key implication of the analysis is that rigorous studies with selection on unobservables can provide unbiased estimates where randomisation is not feasible or ethical. This includes analysis of existing survey or administrative data using natural experimental approaches, which are an underutilised approach in WASH impact evaluation. Ranked by expected mean squared error, the most accurate findings, relative to the benchmark estimate, were from RDD, credible IV, and methods incorporating baseline outcomes in estimation through DD or matching. Matching on local comparisons, nearest neighbour matching, and matching using rich controls followed. The strongest evidence for accuracy is for regression discontinuity design, which across 173 separate estimates from four studies, was able to remove 94 percent of bias on average, with expected MSE equal to 0.0004. Double difference estimation was able to remove 56 percent of bias with expected MSE equal to 0.02, although there were only two estimates from a single study available. However, matching on baseline outcome, which is similar to DD, also on average removed 56 percent of bias with expected MSE of 0.004, across 15 estimates. In contrast, OLS only removed 34 percent of bias with expected MSE equal to 0.18.

As predicted by theory, the accuracy of some estimators was dependent on effective implementation. One estimate suggested IV estimation reduced error by 95 percent with expected MSE less than 0.000, when the instrument

satisfied the exclusion restriction and was strongly correlated with outcome. But matching on a weak instrument can produce a less consistent estimate with more bias than OLS. For matching estimators, the most important characteristic was the use of 'rich controls', leading to 83 percent bias reduction on average across 116 estimates, although with relatively high expected MSE of 0.15. Nearest neighbour matching also outperformed other matching methods, accounting for 52 percent of bias with expected MSE of 0.07.

The findings confirmed some of the decision rules incorporated in the critical appraisal tool in Chapter 4 (Appendix A), such as on the use of baseline covariates or on exogeneity of instrumental variables. The tool was also updated to incorporate questions relating to matching in the analysis of the relationship between WASH provision and mortality in Chapter 6. This included assessment of whether NRS used baseline outcomes, baseline covariates at household and community levels, geographically local matches, and whether outcomes were measured by long recall.

A final comment is warranted about the generalisability of the findings, given the relatively small number of internal replication studies that exist on international development topics. Firstly, the interventions are restricted largely to conditional cash transfers, an approach which has been extensively tested using cluster-randomisation. With the exception of the studies in Cambodia, Tanzania and Tonga, most evidence from internal replications is from Latin America. There may therefore be legitimate concerns about transferability of the evidence to other contexts and sectors, including WASH where no internal replication studies have yet been done according to the searches here.

Chapter 6 Why water supply, sanitation and hygiene are essential for global health

> "[H]ygiene behaviour appears to be universal in human beings, and driven by factors other than wanting to avoid disease. As African mothers told us 'everybody wants to be clean'. Nobody likes dirt as it is unattractive, disgusting and stigmatizing."
>
> Curtis (2001, p.76)

6.1 Introduction

This Chapter examines the relationship between WASH access and child diarrhoea death to address Thesis Question 4: what are the effects of WASH provision on child mortality and do the effects vary by intervention and technology? The analysis was motivated by the lack of existing systematic evidence on childhood survival attributable to improved water, sanitation and hygiene, despite the large numbers of studies reporting mortality (Chapter 3, Section 3.4.1). It also provides another opportunity to further test and refine the risk-of-bias approach developed in this Thesis, by examining the relationship between probable biases, as identified in the tool, and the empirical evidence of bias for an outcome variable which, unlike reported infection, is thought not to be subject to serious biases in reporting.

Section 6.2 presents an overview of the policy and research issues in analysing the relationship between WASH and mortality in childhood. Section 6.3 reviews the existence systematic review evidence on WASH and diarrhoea. Section 6.4 presents search and inclusion decisions for the systematic review and data collection for statistical meta-analysis. Section 6.5 critically appraises the included studies. Section 6.6 presents the meta-analysis results. Section 6.7 concludes and presents revised estimates of the global burden of disease due to inadequate WASH.

6.2 Policy and research issues in estimating the impact of WASH on mortality

How fundamentally important are water, sanitation and hygiene for human life, health and happiness? The psychologist Abraham Maslow (1943) proposed a hierarchy of goals for human life in the following order: "physiological, safety, love, 'esteem, and self-actualization" basic needs. Often referred to as a pyramid (though not specifically by Maslow), the physiological needs at the pyramid's base relate to 'homeostasis' or healthy regulation of the human body's metabolism via sufficient access to air, water, nutrition, warmth, rest (including sleep) and the means to excrete. Maslow placed safety needs just above physiological needs, which he linked specifically to safety from illness and pain in childhood, as well as from 'wild animals' and 'assault' throughout the life-course.[136] It is quite difficult to over-emphasise the contribution of sufficient water, sanitation and hygiene to basic needs.

The quote from Val Curtis at the start of this chapter indicates the substantial interest in hygiene from the bottom up, among potential service end-users. There is also great interest from the top down in the impacts of WASH on child mortality in policy communities. This is in part due to the method of calculation of disability-adjusted life years (DALYs) (Cairncross and Valdmanis, 2006), the preferred technical approach to allocating health budgets. For example, the estimated DALYs due to water-related infection in a population are calculated as:

$$DALY = YLL \quad + YLD \quad = \sum_{i}^{N} D_i . L_i + \sum_{i}^{N} I_i . W^D . L_i \qquad (6.1)$$

where *YLL* is years of life lost (per 100,000), equal to the summation over *N* age groups of the number of deaths D_i (per 100,000) in the population due to water-related infection in each age group *i* multiplied by life expectancy at age of death L_i, and *YLD* is years lived with disability, equal to the summation of the number of incidence cases of water-related infection I_i (per 100,000) in each age-group multiplied by the weight given to disability caused by water-related infection W^D and life expectancy (Prüss-Üstün et al., 2003). Every death attributed to infection, especially among children, is therefore

[136] See also Tanner (1995).

weighted heavily in the DALY calculation. In contrast, a calculation of YLD based on numbers of days experiencing diarrhoeal disease is rather smaller in endemic circumstances, since the typical child diarrhoeal risk among populations lacking access to clean drinking water may be three episodes per year (Clasen et al., 2015). For example, the recent global burden of disease (GBD) exercise estimates YLL for acute lower-respiratory tract infections at over 1,300 deaths per 100,000 and diarrhoea at 960 deaths per 100,000 (GBD 2016 Cause of Death Collaborators, 2017a). These are the third and fourth highest numbers of years of life lost to a single disease among all causes of mortality (and the highest among communicable diseases). In contrast, years lived with disability were estimated at one-tenth of the level of YLLs for diarrhoea (100 per 100,000) and around 1 percent (10 per 100,000) of YLLs for lower-respiratory tract infections (GBD, 2017b).[137]

Churchill et al. (1987) were pessimistic about the potential for water and sanitation projects alone to improve health, but instead argued persuasively that improved water supply could be justified by the substantial economic value of the time-savings in water collection, the opportunity costs of which had already been studied by Cairncross and Cliff (1987), and more extensively since (e.g., Sorenson et al., 2011). It is worth quoting Churchill et al. in full on the health benefits of WASH:

> "The available evidence suggests that there is a very tenuous link between improvements in health and investments in water supply and sanitation services. The best that can be said is that these services may be necessary, but not sufficient, to achieve any tangible effects on morbidity and mortality. The complex chain through which disease is transmitted does not lend itself to simple interventions. Human behavior and its interaction with the environment are just as important in determining overall health status as availability of clean water. Improvements in health are highly correlated with literacy, level of female education, and income, rather than the level of water

[137] In addition, road injuries caused the fifth biggest numbers of YLL at 817 per 100,000 (of which pedestrian road injuries contributed 290 per 100,000) and were in the top 20 causes of YLD at around 200 per 100,000 (pedestrian injuries contributing one-quarter of these). While musculoskeletal disorders caused 31 YLL per 100,000, lower back and neck pain was the biggest single cause of YLD (over 1,000 per 100,000) and other musculoskeletal disorders were the seventh highest (over 500 YLD per 100,000). Animal contact was estimated to contribute 58 per 100,000 YLL and around 30 YLD per 100,000.

> and sanitation services. Thus, in practice, human behavior, particularly in low-income rural areas, has overwhelmed any theoretical links between improved services and improved health."
>
> Churchill et al. (1987, p.ix)

There can be little reason to doubt the value of income or education, particularly of children's carers who are usually women, in improving child health.[138] Furthermore, there has been an explosion in the production of studies that are able to link WASH provision with health, and a large number of syntheses of these studies (Chapter 3, Section 3.4.2). The most common outcome indicator collected in health impact evaluations, and synthesised in systematic reviews, is diarrhoea morbidity. It is presumably collected as a proxy for diarrhoea mortality, since it is easier to measure for financial and ethical reasons (Briscoe et al., 1985). But it may be a poor proxy for diarrhoea mortality due to censoring of data, particularly in observational studies and cluster-RCTs where recruitment of individuals is done after randomisation, or in studies (including RCTs) where children of different ages, and therefore lengths of exposure, are followed up concurrently. Furthermore, diarrhoeal disease prevalence – number of days with diarrhoea over a period – is thought to be more closely correlated with mortality than diarrhoea incidence – number of distinct diarrhoea spells over a period (Morris et al., 1996; Schmidt et al., 2011).

It may also be the case that improved water supply needs less behaviour change programming than other WASH technologies.[139] Water supply may also be an enabling factor for basic sanitation and hygiene if people do not like to use latrines without water availability and/or are unable to wash their hands with soap. For example, a survey was undertaken in 1956 by the *Serviço Especial de Saúde Pública* of Brazil in Palmares, a town in the north-east of the country, to examine the correlation between diarrhoeal mortality among infant and water supply source. The findings indicated that mortality rates among infants living in dwellings where water was collected from outside were approximately triple those in households with water

[138] For a meta-analysis of observational studies on this, see Charmarbagwala et al. (2004).
[139] Although this may not always be the case, if an existing, unimproved water supply provides other individual or community needs such as the ability to socialise, as noted in Chapter 1 (Figure 1.6).

connections (Table 6.1). Furthermore, mortality appeared unchanged whether the source of water was a public faucet or an unprotected well.

Table 6.1 Diarrhoea deaths in urban Brazil

Type of water supply	*Percent of deaths among infants <4 mos.*
Public water system	
House connection	20.0
Outside faucet < 100 m from dwelling	57.1
Outside faucet > 100 m from dwelling	68.0
Outside, unprotected well	57.6

Source: Wagner and Lanoix (1959, p.18).

Wagner and Lanoix (1959) provided two interpretations for the findings. Firstly, they suggested that the reason for the same mortality rate between outside water from the public system and unprotected wells was that public faucet water is re-contaminated between source and point-of-use. Secondly, they stated that "[w]hen water is available and conveniently reached by people, the tendency is to use it in abundant quantities, as a result of which personal cleanliness is maintained. Public health officials have believed for some time that the health benefits deriving from the construction of water-supply systems are considerably reduced unless water is made readily available not only for drinking purposes but also for domestic use and the improvement of personal hygiene" (p.17).[140]

Another study of a World Bank (1998) piped water project providing household connections in rural Paraguay, found that the risk of under-5 mortality in communities without piped water systems was 7.4 times higher than that in communities with the piped water systems. Furthermore, once household connections were completed, the death rate dropped even further to virtually nil (risk ratio=20.5). Although the study indicates that "[c]limate, topography, and water access were similar" World Bank (1998, p.23), neither this, nor the example above from Brazil, formally controlled for possible

[140] It is worth noting that child weaning started early in this part of north-eastern Brazil. One would not usually expect an impact on mortality of improved WASH among infants who are breast-feeding (Gautam et al., 2017).

confounding of the relationship between water supply and mortality, as noted by both study authors.[141]

A useful starting point for analysis of the effect of water, sanitation and hygiene on child survival is Mosley and Chen (1984). This framework links child survival to proximate factors such as nutrient intake, use of health services, childcare practices, which in turn depend on underlying biological factors at the level of the child (e.g., sex, age of mother, birth order, and birth interval), household behavioural and socioeconomic factors (e.g., childcare practices such as breastfeeding, water, sanitation and hygiene behaviour, aspects of housing quality like floor material, and household income), and environmental factors including service provision (e.g., community water, sanitation and hygiene, health care). Water, sanitation and hygiene feature at both household and community levels, reflecting the private and public domains of transmission of infectious disease, as well as the availability of WASH services at community level such as public toilets and health facilities (Cairncross et al., 1996). In the health production function literature, survival is modelled as the product of decision-making, accounting for child, household and environmental factors (e.g., Charmarbagwala et al., 2004):

$$S_i = f(C_i, H_i, E_i) \qquad (6.2)$$

where S_i is the survival status of child i, and C_i, H_i, and E_i represent child-level, household-level and environmental determinants, respectively. Household decisions about health are taken jointly with decisions about household member's time allocation and budget, hence many factors may either be caused by health status, such as health care seeking and aversion behaviours like hygiene practices (reverse-causality), or determined simultaneously by other, possibly unobserved factors (Rosenzweig and Schultz, 1983). Unobservables affecting child survival prospects in households (termed 'frailty' effects) include genetic factors and family-specific behaviours such as son preference (Sen, 1998) and attitudes to childcare (e.g., Heckman and Singer, 1984).

[141] The data from these studies were not admissible for inclusion in meta-analysis because population figures were not given, from which standard errors could be calculated.

Other sources of confounding of the relationship between WASH and mortality include confounding by cause of death, where deaths caused by factors not related to infectious disease are included in the mortality estimates; confounding by age (Blum and Feachem, 1983), where crude deaths of groups of different ages, and hence susceptibility to infectious diseases, are compared; differences due to context, such as where short duration of breastfeeding leads to increased susceptibility of infants to diarrhoea (Gamper-Rabindran et al., 2008); confounding by WASH intervention measure (full subsidy, promotion or exposure variable); the absolute position of the WASH improvement on the water, sanitation or hygiene ladders and the position relative to the previous position on the ladder (i.e., relative to baseline water and sanitation access) (Fewtrell and Colford, 2004); and confounding by co-interventions, such as where areas with piped water or sanitation are likely to have access to other health inputs affecting mortality (e.g., health care, nutrition supplements, public health infrastructure) (Jalan and Ravallion, 2003). On the other hand, evidence suggests that confounding bias due to self-reporting is not problematic for all-cause mortality, and less problematic for cause-specific mortality (Wood et al., 2008, Savović et al., 2012). However, it is also thought that poor people are likely to over-report use of WASH technologies and underreport disease (Briscoe et al., 1985), including cause-specific deaths obtained using 'verbal autopsy' in carer surveys and through vital registration (Anker, 1997; Victora et al., 2001).

Analysis of the causal relationship between water, sanitation and hygiene and survival, needs to account for these sources of confounding as far as possible. One approach to resolve the problem uses experimental design – randomised assignment of WASH hardware and/or promotional approaches. Randomisation of intervention, across a sufficiently large sample, ensures temporality (cause precedes effect – as noted in Chapter 4, Table 4.4) and should balance frailty effects between treatment arms. While analyses of the causal effect of WASH on child mortality are available from experimental studies, as shown in Section 6.4 below, most analyses use non-randomised designs, many based on covariate-adjusted analysis. This is partly because it is unethical to let people die in the course of intervention research when oral rehydration salts (ORS) or medical treatment may be easily provided to severely ill children (e.g., Briscoe et al., 1985; Daniels et al., 1990b). Furthermore, the sample requirements to estimate effects on

mortality with statistical precision are usually beyond what is affordable or feasible in single studies.

Therefore, the second main approach is to use modelling of observational data. Where survey data at the level of the individual child is used, for children of different ages, not all of whom will have reached the upper age cut-off for the mortality rate in question (e.g., age 5 in the case of under-5 mortality), survival models are applicable such as the proportional hazards model (e.g., Masset and White, 2003).[142] Where the outcome being measured is number of events per person over a specified time period, alternative approaches can be used, including those discussed in Chapter 4, Sections 4.3 and 4.4, such as double differences, statistical matching and adjusted regression models. For example, a retrospective method common to epidemiology which uses statistical matching alongside adjusted logistic regression is the case-control design (Briscoe et al., 1985). These methods are appropriate where data are dropped for children who have not yet completed the age cut-off for the mortality rate (i.e., censored observations). For example, for neonates (children aged below 1 month) it is possible to drop observations on children born in the month of the interview and estimate using logistic regression (Masset and White, 2003).

6.3 Existing review evidence

There is a big systematic review literature examining the effects of water, sanitation and hygiene technologies on diarrhoeal disease in L&MICs. As noted in Chapter 3, Section 3.4.2, the earliest reviews covered faeces-related infections associated with water and sanitation provision including diarrhoea (Esrey et al., 1985) and water-related infections (Esrey et al., 1991). Esrey (1991) concluded that "safe excreta disposal and proper use of water for personal and domestic hygiene appear to be more important than drinking water quality in achieving broad health impacts" (p.31).

[142] The proportional hazards model assumes that the risk of death for any age can be calculated by adjusting the baseline risk by an exponentiated set of factors: $h_i(t) = h_0(t)e^{\beta_1 C_i + \beta_2 H_i + \beta_3 E_i + \varepsilon_i}$, where $h_i(t)$ represents the mortality rate at time (t) for individual i and ho(t) is the age-specific baseline hazard, which is the mortality risk at each age in the case where all explanatory variables are equal to zero. It is therefore similar to the constant term in a standard regression model. The βs refer to the estimated coefficient parameters on child, household and environmental characteristics and ε_i is the error term incorporating unobservable frailty.

Fewtrell and Colford (2004; Fewtrell et al., 2005) meta-analysed 60 studies, finding that both hygiene education and household water treatment reduced the risk of diarrhoea disease by about 40 per cent each in L&MICs, while sanitation provision or water supply reduced the risk by only around 20 per cent each. A meta-analysis of 33 studies conducted by Clasen et al. (2006, updated in 2015) also supported the finding that water treatment at POU, particularly filtration, is more effective in reducing diarrhoea risk than other types of water improvements. These findings were replicated in Hunter (2009) and the WHO (Wolf et al., 2014, 2018). Interventions appeared to be more effective when a safe water storage container was also provided (Clasen et al., 2015), as it is for example in filtration devices from which water is accessed through a tap.

A few meta-analyses of higher quality studies found that piped water to households significantly reduced diarrhoea morbidity (Waddington et al., 2009; Wolf et al., 2018). Wolf et al. (2018) also defined piped water according to reliability and quality, finding big impacts, although small numbers of studies.

The evidence on sanitation is mixed. Firstly, until the last decade there were few impact evaluations of sanitation impact covering more than a small number of clusters. Secondly, previous reviews did not take clustering into account. Thus, earlier reviews estimated between 25 and 35 percent reductions in diarrhoea from sanitation (Fewtrell and Colford, 2004; Waddington et al., 2009; Norman et al., 2010; Wolf et al., 2014, 2018). Replacing on-site sanitation with water-based sewerage was estimated to reduce the incidence of diarrhoea by around 30 percent, though it may not always be a suitable solution given the maintenance costs (Norman et al., 2010). The review for Cochrane by Clasen et al. (2010) did not conduct meta-analysis because none of the studies at that point had taken clustering of observations into account in calculating standard errors. However, it also omitted a quasi-experiment conducted of a city-wide sanitation programme in Brazil, which collected longitudinal cohort data before and after intervention, interpretable as interrupted-time series design (Barreto et al., 2007), together with extensive mediator analysis (Genser et al., 2008; Barreto et al., 2010).

Meta-analyses suggested hand hygiene reduced reported diarrhoea morbidity by between 30 and 50 percent (Curtis and Cairncross, 2003; Aiello et al., 2008; Waddington et al., 2009; Cairncross et al., 2010; Ejemot-Nwadiaro et al., 2015; Wolf et al., 2018). Soap provision appeared to be particularly effective (Aiello et al., 2009; Waddington et al., 2009). A question about the effectiveness of hand hygiene in contexts with limited water supply, which would limit study participants' abilities to practice domestic hygiene, suggested that studies with below average effects on diarrhoea infection in the meta-analysis of hand hygiene by Curtis and Cairncross (2003) were indeed done where water supply availability was limited.[143] In Lima, vendors sold water from tanker trucks on the street corner, for 10 to 15 times the price paid by those with house connections (Yeager et al., 1991). In Malawi, the context was a refugee camp for Mozambican refugees, where water supplies were likely to be insufficient (Peterson et al., 1998). In Burundi, mean consumption of water was 6 litres per capita per day (Birmingham et al., 1997). Furthermore, in a study with null findings in Bangladesh (Hoque et al., 1999), hand hygiene included reported "ash, or soil for handwashing after defaecation" (Curtis and Cairncross; p.277), methods which are not commonly recognised as improved practices (Chapter 1 Table 1.1). Other reviews of the effects of handwashing on respiratory illness found 20 percent reduction on average (Rabie and Curtis, 2006; Aiello et al., 2008; Mbakaya et al., 2017), but most of the evidence was from high income countries.

A common finding from meta-analysis of indirect study comparisons (that is, findings across different contexts) is that bundling WASH together does not produce additional effects in comparison with single water, sanitation or hygiene technologies (Fewtrell and Colford, 2004). For example, White and Gunnarson (2008: 17) concluded that "the health impact of combined methods has not been found to be stronger than any single approaches" (p.17). There are two main reasons why the provision of multiple WASH may not lead to bigger observed effects on reported diarrhoea than single interventions. First, the starting conditions differ, and distance moved up the WASH ladder is likely to be correlated with reduction in disease, a factor that has been explicitly modelled in network meta-analysis Wolf et al. (2014, 2018). The second reason is reporting bias, due to participants becoming fatigued after repeated measurements, as discussed below.

[143] Sandy Cairncross, pers. comm.

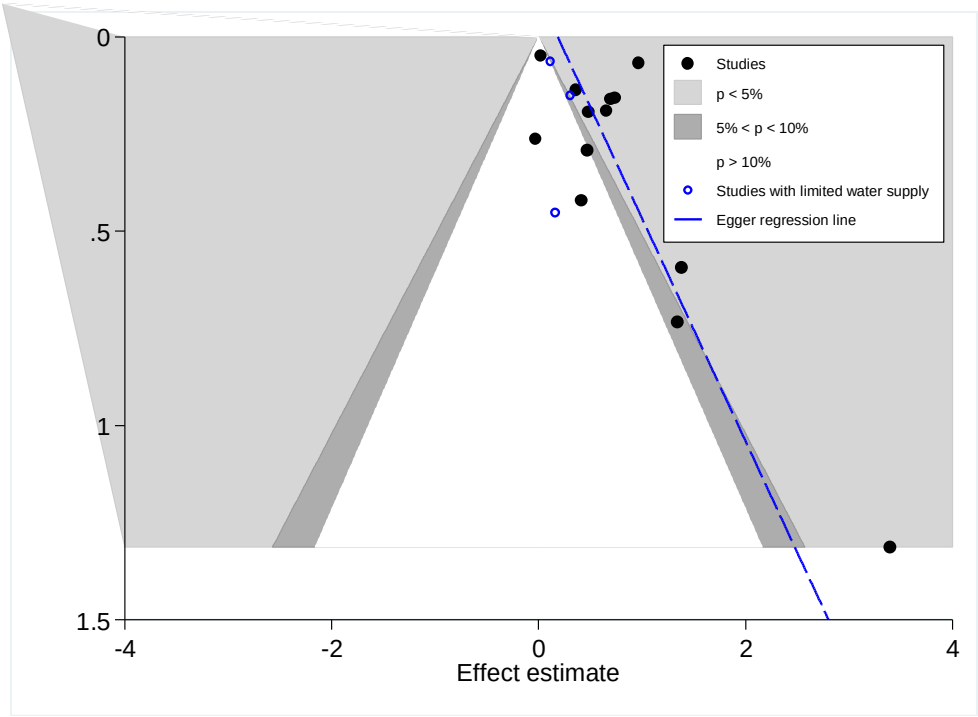
The large number of systematic reviews focusing on diarrhoea impacts of water quality and hygiene improvements is an area where sufficient studies may exist for meaningful analysis of bias. Two forms of bias that have been evaluated in depth are publication bias and biases due to lack of blinding of participants and observers. For example, while often showing strong impacts on reported disease, much of the evidence on water quality and hygiene comes from trials conducted at zero or negligible cost to participants, with frequent within-intervention follow-up and possibilities for bias, and over relatively short periods of time and small samples of beneficiaries. The findings may therefore be superficial in their applicability to WASH policy and programming. There are also concerns about conflict of interest, possibly leading to publication biases, due to many of these studies being funded by private manufacturers (Waddington et al., 2009).

Curtis and Cairncross (2003) were the first WASH meta-analysts to test formally for publication bias in diarrhoea morbidity estimates, reporting Begg-Mazumdar (1994) test ($p>0.11$; 17 studies) and suggesting this was not sufficient statistical evidence for publication bias. However, a funnel graph was not reported. It is recommended that statistical analysis of small-study effects is done alongside visual graphical analysis in diagnosis (Higgins and Green, 2011). Reconstructing the analysis from reported data (Figure 6.1), using contour-enhanced funnel graph estimation (Peters et al., 2008), suggested possible asymmetry in the plot in the areas of statistical significance, although the test statistic is not conclusive (Egger et al., 1997, test $p>0.16$). The idea behind contour-enhanced funnels is that they can account for factors confounding the relationship between effect size and standard error, which are not related to publication bias. One such factor is bias in the impact estimate, which is expected to increase the magnitude of the effect estimate in NRS, all else equal, as shown in Chapter 4, Section 4.2. This would be reflected in missing small-sample studies in the area of non-statistical significance at the bottom of Figure 6.1, as indicated. Therefore, the asymmetry in this case may not be related to publication bias but rather to the inclusion in the review of studies with 'high risk of bias', including self-selected treatment groups.

Fewtrell and Colford (2004) found that there may be evidence for publication bias in WASH studies in L&MICs, especially in studies of water treatment

(Begg-Mazumdar, 1994, test p<0.1; 15 studies). Subsequently, Clasen et al. (2006) presented an asymmetric funnel plot for water treatment studies, suggesting small sample studies with smaller effect sizes may be being suppressed. However, they noted that "[w]e chose not to present results from statistical analysis of publication bias" (p.6), and "[s]ince we found substantial evidence of [clinical and methodological] heterogeneity, we cannot conclude that the funnel plot demonstrates evidence of publication bias in this case" (p.12). Wolf et al. (2018, p.519) also stated "[t]here was no evidence of funnel-plot asymmetry and small study effects in any of the WaSH meta-analyses" included in that review. This is a surprising finding, since publication bias has been shown to exist in literatures from all disciplines (Rothstein et al., 2005). Further examination of the funnel graphs indicated Wolf et al. (2018) did not use methods of small-study analysis which take account of other sources of funnel graph asymmetry, such as bias in effect estimation due to inclusion of broad study designs in meta-analysis (Peters et al., 2008), as observed above.

Figure 6.1 Funnel graph with small-study effects regression line



Note: effect estimate shows protective effect of hygiene on morbidity.
Source: author using data reported in Curtis and Cairncross (2003).

As noted in Chapter 4, Section 4.4, there are concerns about effectiveness of WASH interventions in reducing morbidity due to concerns about the quality of self- and carer-reported health outcomes, particularly where survey

participants are exposed to repeated measurement in open (unblinded) trials (Schmidt and Cairncross, 2009; Zwane et al., 2011). One advantage of water treatment technology with respect to conducting trials is that it is possible to blind participants – for example, by providing plastic bottles but no instructions to put them in direct sunlight for ultraviolet (UV) filtration (Conroy et al., 1996). Schmidt and Cairncross (2009) famously reported that blinded studies of household water treatment estimated impacts that were not significantly different from zero (RR=0.91; 95%CI=0.82, 1.02; evidence from 3 studies pooled by author).[144] Other reviews of household water treatment and storage trials found smaller or null effects once double blinding was taken into account (Clasen et al., 2006; Waddington et al., 2009; Hunter, 2009; Clasen et al., 2015; Wolf et al., 2018) (Table 6.2).

Others noted that water treatment technologies were more effective where adherence was higher (Arnold and Colford, 2007; Waddington et al., 2009; Clasen et al., 2015). One review found that "water quality interventions conducted over longer periods tend to show smaller effectiveness, while compliance rates, and therefore impacts, appear to fall markedly over time" (Waddington et al., 2009; iii). As noted in Chapter 1, Section 1.4.2, it appears difficult to encourage children's carers to change behaviour when the main benefits of a new technology, such as reducing a child's disease rate, are hard for them to observe. Schmidt and Cairncross (2009) concluded that "widespread promotion of household water treatment is premature given the available evidence" (p. 986). There is therefore considerable controversy as to the role and scalability of water treatment in combating diarrhoeal disease.

Issues affecting the quality of self-reported diarrhoea morbidity may also affect hygiene evaluations. Although no studies with double blinding of participants and outcome assessors have been conducted of hygiene interventions in L&MICs, blinding of outcome assessors is achievable, for example where participants were provided children's reading material unrelated to hygiene (Luby et al., 2006). One systematic review found a smaller, but still statistically significant, impact of hand hygiene on diarrhoeal morbidity in blinded trials (RR=0.80, 95%CI=0.67, 0.94; 4 studies) (Ejemot-Nwadiaro et al., 2015).

[144] Schmidt and Cairncross (2009) reported the protective effect of household water treatment on reported morbidity. This was inverted for comparability with other calculations in this Thesis.

It appears to be increasingly common to adjust for lack of blinding using Bayesian methods (Table 6.2). Hunter (2009) was the first to propose a bias correction procedure to water treatment studies drawing on bias coefficients from between-study meta-epidemiology findings (Wood et al., 2008). In the updated Cochrane drinking water treatment review by Clasen et al. (2015), similar bias correction factors were also applied, although the authors noted that "we urge caution in relying on these adjusted estimates since the basis for the adjustment is from clinical (mainly drug) studies that may not be transferable to field studies of environmental interventions" (p.9). Wolf et al. (2018) also adjusted the effects of household water treatment and hygiene for bias due to lack of blinding, but not water supply and sanitation, arguing that water supply and sanitation have recognised benefits over and above health impacts, whereas water treatment and hygiene "usually aim exclusively to improve health which is apparent to the recipient" (p.512). The correction factor for hygiene studies was particularly large, yielding a highly imprecise estimate (OR=0.90, 95%CI=0.37, 2.17; 33 studies) that was much bigger than the bias from single blinding estimated in the systematic review of RCTs by Ejemot-Nwadiaro et al. (2015). In addition, these adjustments would not be appropriate for any observational studies which were not conducted under trial conditions, or in clustered trials where informed consent did not mention an intervention (Schmidt, 2014). In such circumstances, respondents would not associate measurement with the intervention, so arguably having fewer incentives to misreport.

A few other relevant reviews incorporated estimates of mortality reduction due to factors associated with WASH provision. Morris et al. (2003) reviewed evidence on studies reporting cause-related mortality among under-5s, estimating 22 percent of deaths were due to diarrhoea and 20 percent to pneumonia. Benova et al. (2014) estimated substantial reductions in maternal mortality due to improvements in sanitation (OR=0.32, 95%CI=0.20, 0.51) and water access (OR=0.57, 95%CI=0.39, 0.83). Re-analysis of the data in Benova et al. (2014) suggested these findings were driven by improvements in water supply but not water quality.[145]

[145] Water supply pooled effect OR=0.42 (95%CI=0.29, 0.83, I-squared=0%, evidence from 2 studies). Water treatment pooled effect OR=0.75 (95%CI=0.49, 1.14, I-squared=24%, evidence from 2 studies).

Table 6.2 Bias adjustment in meta-analyses of diarrhoea morbidity

	Confidence rating	*Pooled effect*	*95% CI*		*I²*	*# studies*	*Bias ratio**	*Comments*
Trials with blinding of participants and/or outcome assessors								
Clasen et al. (2006)	High	1.07	0.88	1.30	0%	2	NA	POU water treatment; double blind studies
Waddington et al. (2009)	Medium	0.76	0.59	0.97	NR	4	1.13	Handwashing with soap and health education; single blind studies
Cairncross et al. (2010)	Medium	0.93	0.70	1.33	NR	3	1.60	POU water treatment; double blind studies
Ejemot-Nwadiaro et al. (2015)	High	0.80	0.67	0.94	71%	4	1.29	Handwashing in community; single blind studies
Clasen et al. (2015)	Moderate	1.07	0.97	1.17	0%	4	1.57	Chlorination; double blind studies; high quality evidence
Clasen et al. (2015)^	Moderate	0.80	0.68	0.94	20%	5	1.95	Filtration; double blind studies; high quality
Bayesian meta-analysis with error correction								
Hunter (2009)	Medium	0.85	0.76	0.86	NA	28	1.52	POU water treatment
Clasen et al. (2015)	Medium	0.70	0.64	0.77	NA	55	1.25	All water treatment
Clasen et al. (2015)	Medium	0.65	0.40	1.09	NA	7	1.35	Flocculation and disinfection
Clasen et al. (2015)	Medium	0.80	0.69	0.92	NA	19	1.11	Chlorination
Wolf et al. (2018)	Low	0.91	0.70	1.18	NA	18	1.20	Chlorination

	Confidence rating	*Pooled effect*	*95% CI*		*I²*	*# studies*	*Bias ratio**	*Comments*
Clasen et al. (2015)	Medium	0.62	0.55	0.70	NA	23	1.29	Filtration
Wolf et al. (2018)	Low	0.60	0.42	0.84	NA	15	1.22	Filtration
Wolf et al. (2018)	Low	0.52	0.35	0.77	NA	8	1.33	Filtration with safe storage
Clasen et al. (2015)	Medium	0.80	0.60	1.01	NA	6	1.18	SODIS
Wolf et al. (2018)	Low	0.88	0.60	1.27	NA	5	1.31	SODIS
Wolf et al. (2018)	Low	0.90	0.37	2.17	NA	33	1.29	Handwashing

Notes: * author's calculation by dividing 'corrected' effect size by 'uncorrected' effect size; ^ includes evidence from low-income contexts in high income countries; NR not reported; NA not applicable; confidence ratings taken from census of WASH studies (Chapter 3, Section 3.4.2) are reported in full in Chirgwin et al. (2021).

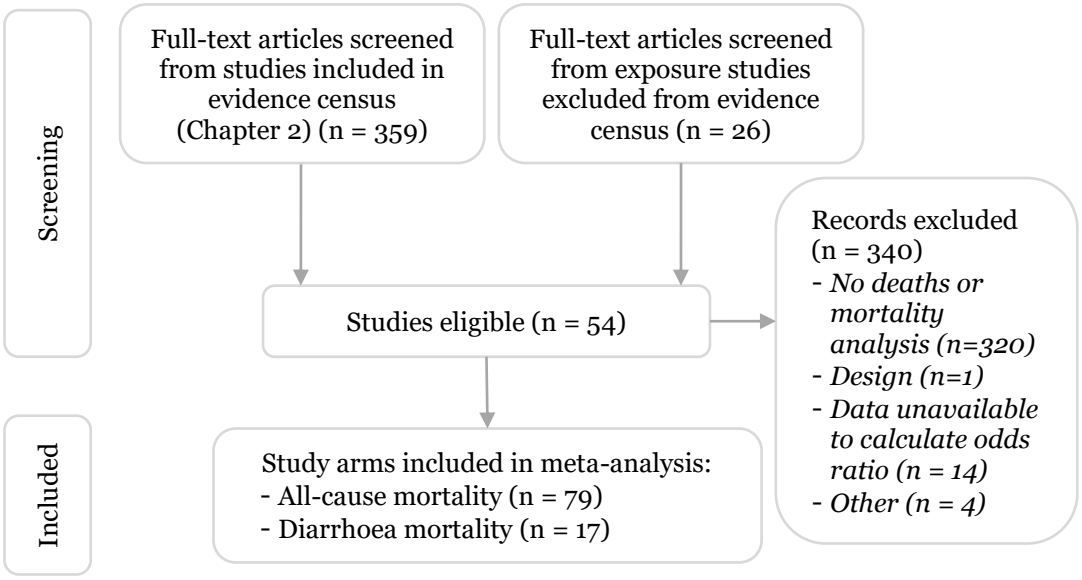
Charmarbagwala et al. (2004) meta-analysed observational studies examining the association between child mortality and water and sanitation. Of 15 studies that had incorporated water and sanitation variables in regression analysis of infant and child survival, they found that water supply and sanitation were strongly associated with infant mortality, but only water supply seemed to be associated with lower child mortality.[146] However, Charmarbagwala et al. (2004) analysed t-statistics which are a noisy indicator of effect as they are dependent on both the size of the effect and the study sample size. White and Gunnarson (2008) incorporated mortality studies in a systematic review of WASH impacts. Although the authors did report risk ratios from meta-analyses included in that review, the summary of findings in that review used the 'voting method' (Smith and Glass, 1977). The preferred approach is to use an effect size (e.g., an odds ratio or risk difference) that reflects the magnitude of the effect, and to incorporate the study sample size in the weighting scheme in calculating the pooled effect across studies (e.g., Smith and Glass, 1977; Waddington et al., 2012).

6.4 Data collection

The protocol for this systematic review was registered with Prospero and the Campbell Collaboration (Waddington and Cairncross, 2020).[147] The author started from the studies reporting all-cause and diarrhoea-related mortality in the census of studies in the WASH evidence map (summarised in Chapter 3, Section 3.4.1). However, the census reported mortality where the studies analysed mortality rates between groups. Therefore, not all included studies that reported deaths, for example in the participant flow diagram, were coded under mortality outcomes in the map itself (e.g., Boisson et al., 2010). Therefore, the author re-reviewed the participant flow diagrams in all studies to obtain crude mortality rates for field trials by intervention group. In addition, screening was done of studies that were excluded from the evidence map because they were not linked to a particular intervention. The search process is detailed in Figure 6.2.

[146] They also found the converse for child nutrition, that household sanitation access appeared to be more important than water supply. There was no effect of community (shared) sanitation on nutrition, but community water availability was associated with better nutrition, albeit less strongly than household connections.
[147] 'Water, sanitation and hygiene for reducing death in children in low- and middle-income countries', CRD42020210694. Available at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=210694 (accessed 30 November 2020).

Figure 6.2 Study search flow



A large number of longitudinal studies (whether randomised or non-randomised) did not include a participant flow diagram, or information with which to reconstruct it. In particular, almost no study published in a social science journal or working paper provided full details on the participant flow from recruitment through to follow-ups by intervention group. Some papers indicated that they had collected mortality information but did not report it, for example because the rarity of the events limited statistical power (Duflo et al., 2015), or because mortality was to be examined in forthcoming publications (Sinharoy et al., 2017). Some reported the total number of deaths recorded, not deaths by intervention group (Tonglet et al., 1992; Hunter et al., 2013; Kirby et al., 2017). Others simply indicated that child mortality rates were similar in intervention and control groups (Stanton et al., 1988).

Many studies providing participant flow diagrams did not give the reasons for losses to follow-up, from which all-cause crude mortality rates could be derived. Crude mortality rates by intervention group were not reported in some studies (e.g., Boisson et al., 2013), and in other cases have not been published yet although the study protocol indicated these data would become available (Brown et al., 2015).[148] In order to obtain relevant effect sizes from

[148] Where deaths were reported but were equal to zero in any group, conventional practice was followed by adding 0.5 to all cells of the two-by-two frequency table (Cole et al., 2012; Gyorkos et al., 2013; Luby et al., 2006; Mengistie et al., 2013; Semenza et al., 1998). One study with 5-months of follow-up reported zero deaths

all non-randomised studies examining the relationship between WASH and mortality. One paper included in the WASH evidence map was excluded from analysis because it simulated mortality in the control group (Meddings et al., 2004). Another study from Mexico was not includable in analysis because it used state-level baseline diarrhoea mortality data to estimate effects on nutrition outcomes and schooling outcomes (Venkataramani and Balhotra, 2013).

In addition, exposure studies excluded from the evidence census (Chapter 3, Section 3.3) and studies included in the meta-analysis by Charmarbagwala et al. (2004) were assessed for inclusion. Two studies were excluded (Da Vanzo et al., 1983; Da Vanzo, 1988) to avoid statistical dependency, as they used the same data source as an included study (Da Vanzo and Habicht, 1986). One did not provide the variable means from which effect sizes could be calculated (e.g., Lee et al., 1997). Another did not provide the regression coefficients for water or sanitation variables (Bicego and Boerma, 1993). One study included the share of households with inadequate water supply (Terra de Souza, 1999) which was not directly comparable with other studies.[149] Two studies included only a composite measure for water and sanitation access (Kishor and Parasuraman, 1998; White et al., 2005). One study that included interaction terms for water and sanitation with maternal literacy and breastfeeding practices, was excluded as the variable means were not presented for the full partial differential to be calculated (Esrey and Habicht, 1988).[150] One study was excluded because it reported only pre-test post-test results for the water and sanitation intervention (Newman et al., 2002). It was not possible to calculate the hazard ratio in a study in Brazil (Sastry, 1996), nor to calculate the odds ratio for studies reporting proportional hazards in Ethiopia (Gebretsadik and Gabreyohannes 2016), India (Masset and White, 2003), Ghana (Gyimah, 2002), Mozambique (Macassa et al.,

in each group and was therefore excluded (Wang et al., 1989). One study was excluded as it did not contain a control group – George et al. (2012) examining arsenic testing and information by a community member versus an external organisation.

[149] Terra de Souza (1999) and Geruso and Spears (2018, 2019) also reported, respectively, the shares of households with inadequate sanitation or openly defaecating. This variable is equivalent to the level of environmental contamination, which was included as an explanatory variable in moderator analysis.

[150] If the equation being estimated is $MR = b_1 WASH + b_2 WASHxLIT + b_3 WASHxFEEDING$, where LIT is maternal literacy and FEEDING is breastfeeding practices, the effect of WASH is calculated as the partial differential at the data means: $\frac{\partial MR}{\partial WASH} = b_1 + b_2\overline{LIT} + b_3\overline{FEEDING}$.

2004) and Senegal (Brockerhoff, 1990), as the baseline risk was not reported. In these cases, the proportional hazard was included in meta-analysis, based on the likelihood of similarity with OR given the low risk of death in the population.

Data were collected from each study on the country, location (rural, urban, nationwide), participant age-group, WASH technology, environmental contamination as represented by community water and sanitation access at baseline, and effect size and standard error using formulae in Appendix C. Baseline water and sanitation were determined by the type that was most frequently used in the control group. Following Fewtrell and Colford (2004), where the study did not report the baseline assessment, the value was imputed for the relevant country, location and year from the Joint Monitoring Programme dataset. In addition, the study design and methods were critically appraised using the risk-of-bias tool (Chapter 4 and Appendix A). It is important to recognise that the risk of bias in these tables refers to the likelihood of bias in the mortality estimate, rather than the overall risk of bias in the study for the other outcomes. To this end, the risk-of-bias tool was slightly modified, in the intervention and outcome domains, as discussed below in Section 6.5. In the end, 54 studies were included, evaluating 87 separate WASH intervention arms. One study was reported in French (Messou et al., 1997a), one was in Portuguese (Rasella, 2003) and one in Spanish (Instituto Apoyo, 2003). The rest were reported in English. All RCTs were published in journals. The summary of included studies is in Table 6.3.

Most RCTs used cluster design, with clustering at the community level; one cluster-RCT pair-matched communities prior to random assignment (Nicholson et al., 2014). Control groups often received standard WASH access with no additional interventions, although occasionally they received another intervention (e.g., all participants received hygiene education in Lule et al., 2005) or a placebo (e.g., Luby et al., 2006, and Bowen et al., 2012, provided children's books, notebooks, pens and pencils to controls). One study used a combination of observational design for the piped water supply versus non-piped comparison groups, and within the comparison arm, prospective random assignment to household water treatment, safe storage and handwashing promotion arm, or control (Semenza et al., 1998).

Most NRS were retrospectively designed, although several used prospective non-randomised controlled designs (Cole et al., 2012; Messou et al., 1997; Rasella, 2003) and several others analysed cohort data (Rhee et al., 2008; Ryder et al., 1985; Semenza et al., 1998). For the retrospective studies, there were two case-controls (Hoque et al., 1999; Victora et al., 1988). One study used pipeline design by enrolling as controls those due to receive the WASH intervention at a later time (Instituto Apoyo, 2000). Several others were able to construct pseudo-panels (repeated cross-section data) from vital registration, census and/or survey data, and applied fixed effects or double-differences regression (Gamper-Rabindran et al., 2008; Rasella, 2003), with more rigorous approaches also incorporating statistical matching of vital registration and/or census data (Galdo and Briceño 2005; Galiani et al., 2005; Granados and Sańchez 2013). A few analysed cross-section survey data using adjusted regression (Fink et al., 2011; Fuentes et al., 2006; Geruso and Spears, 2018) or statistical matching (Abou-Ali et al., 2010).

The mortality rates were computed over a standard period, as mortality measurements will increase over longer exposure periods, all else equal.[151] Gebre et al. (2011; citing Siegel et al., 2004) used the following calculation for CMR_j , the crude mortality rate in study *j* per 1,000 person-years at risk:

$$CMR_j = \frac{D_j}{\frac{t_j}{12}\left(N_j - \frac{D_j + M_j}{2}\right)} \; x\; 1{,}000 \qquad (6.3)$$

where D_j is the number of deaths, t_j is the study follow-up period in months, N_j is the baseline sample size and M_j is the number of people who permanently migrated out of the study area over the follow-up period. This was applied to data collected from included studies. Permanent migrants were usually not reported in included studies, with the exception of Luby et al. (2018).

[151] This is particularly important for comparative measures of mortality rates (effect sizes) that are time sensitive, such as risk differences, but less important for ratio estimates. However, follow-up length was collected from studies and included in meta-regression analysis as it has been shown to be correlated with effect sizes in a previous meta-analysis of diarrhoea morbidity (Waddington et al., 2009).

Table 6.3 Description of studies included in mortality meta-analysis

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Water supply									
Abou-Ali et al. (2010)	Egypt	National	Piped water supply	All-cause mortality	0-59s	Improved	Unimproved	NRS	High risk
Brockerhoff (1990)	Senegal	National	Piped water supply	All-cause mortality	0-15s	Unimproved	Unimproved	NRS	High risk
Brockerhoff and Derose (1996)	Kenya, Madagascar, Malawi, Tanzania, Zambia	National	Piped water supply	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Casterline et al. (1989)	Egypt	National	Piped water supply	All-cause mortality	0-59s	Improved	Unimproved	NRS	High risk
DaVanzo and Habicht (1986)	Malaysia	National	Piped water supply	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Ercumen et al. (2015b)	India	Urban	Continuous piped water supply	All-cause mortality	0-59s	Improved	Improved	NRS	High risk
Fink et al. (2011)	Worldwide	National	Piped water	All-cause mortality	0-59s	N/A	N/A	NRS	High risk
Fuentes et al. (2006)	Cameroon	National	Piped water or covered well	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Fuentes et al. (2006)	Egypt	National	Piped water or covered well	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Fuentes et al. (2006)	Peru	National	Piped water or covered well	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Fuentes et al. (2006)	Uganda	National	Piped water or covered well	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Galiani et al. (2005)	Argentina	Urban	Privatised water supply	All-cause mortality Mortality due to infectious disease and parasites	0-59s	Improved	Unimproved	NRS	Some concerns
Gamper-Rabindran et al. (2008)	Brazil	National	Piped water supply	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Gebretsadik and Gabreyohannes (2016)	Ethiopia	National	Piped water supply	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Geruso and Spears (2018, 2019)	India	National	Piped water supply	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Gyimah (2002)	Ghana	National	Piped water supply	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Hoque et al. (1999)	Bangladesh	Rural	Tube well water storage <2l vs surface water storage >2l	Diarrhoea mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Hoque et al. (1999)	Bangladesh	Rural	Tube well water supply	Mortality due to infectious disease Mortality due to ARI	0-59s	Unimproved	Unimproved	NRS	High risk
Howlader and Bhuiyan (1999)	Bangladesh	National	Piped water supply or public tap	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Kanaiaupuni and Donato (1999)	Mexico	Rural	Source water supply	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Macassa et al. (2004)	Mozambique	National	Piped water supply	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Masset and White (2003)	India	State-wide	Safe water supply	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Mellington and Cameron (1999)	Indonesia	National	Piped water supply	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Ryder et al. (1985)	Panama	Rural	Piped water supply	Diarrhoea mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Semenza et al. (1998)	Uzbekistan	Urban	Piped water supply	Diarrhoea mortality	0-59s	Unimproved	Improved	NRS	High risk
Victora et al. (1988)	Brazil	Urban	Piped water supply	Diarrhoea mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Water treatment									
Boisson et al. (2010)	DRC	Rural	Household water treatment provision (LifeStraw filter)	All-cause mortality	0-59s	Unimproved	Unimproved	RCT	High risk
Conroy et al. (1999)	Kenya	Rural	Solar disinfection (SODIS)	All-cause mortality	0-71s	Unimproved	Unimproved	Quasi-RCT	High risk
Crump et al. (2005)	Kenya	Rural	Household water treatment provision (flocculant)	All-cause mortality	All ages 0-59s	Unimproved	Unimproved	Cluster-RCT	Some concerns
Crump et al. (2005)	Kenya	Rural	Household water treatment provision (chlorine)	All-cause mortality	All ages 0-59s	Unimproved	Unimproved	Cluster-RCT	Some concerns
Du Preez et al. (2011)	Kenya	Rural and urban	Solar disinfection (SODIS)	All-cause mortality	6-59s	Unimproved	Unimproved	RCT	High risk
Ercumen et al. (2015a)	Bangladesh	Rural	Household water treatment (chlorine) and safe storage	All-cause mortality	6-30s	Improved	Unimproved	RCT	Some concerns
Jain et al. (2010)	Ghana	Urban	Household water treatment provision (chlorine)	All-cause mortality	All ages	Unimproved	Unimproved	RCT	High risk
Luby et al. (2006)	Pakistan	Urban	Household water treatment provision (chlorine)	All-cause mortality	All ages	Unimproved	Improved	Cluster-RCT	High risk

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Luby et al. (2006)	Pakistan	Urban	Household water treatment provision (flocculant)	All-cause mortality	All ages	Unimproved	Improved	Cluster-RCT	High risk
Luby et al. (2018)	Bangladesh	Rural	Household water treatment provision (chlorine)	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	Some concerns
Lule et al. (2005)	Uganda	Rural	Household water treatment (chlorine) and safe storage	All-cause mortality	All ages	Unimproved	Unimproved	RCT	High risk
Mengistie et al. (2013)	Ethiopia	Rural	Household water treatment provision (chlorine)	All-cause mortality	0-59s	Unimproved	Unimproved	Cluster-RCT	Some concerns
Morris et al. (2018)	Kenya	Rural	Household water treatment provision (filter)	All-cause mortality	4-16s	Unimproved	Unimproved	RCT	High risk
Null et al. (2018)	Kenya	Rural	Household water treatment provision (chlorine)	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	High risk
Peletz et al. (2012)	Zambia	Rural	Household water treatment (Lifestraw filter) and container	All-cause mortality Diarrhoea mortality	0-23s	Unimproved	Unimproved	RCT	Some concerns
Victora et al. (1988)	Brazil	Urban	Treated piped water supply	Diarrhoea mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Sanitation									
Brockerhoff (1990)	Senegal	National	Latrine access	All-cause mortality	0-15s	Unimproved	Unimproved	NRS	High risk
Casterline et al. (1989)	Egypt	National	Latrine access	All-cause mortality	0-59s	Improved	Unimproved	NRS	High risk

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
DaVanzo and Habicht (1986)	Malaysia	National	Latrine access	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Emerson et al. (2004)	Gambia	Rural	Latrine provision	All-cause mortality	All ages	Unimproved	Unimproved	Cluster-RCT	Some concerns
Fink et al. (2011)	Worldwide	National	Non-open defaecation	All-cause mortality	0-59s	N/A	N/A	NRS	High risk
Fuentes et al. (2006)	Egypt	National	Modern toilet access	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Fuentes et al. (2006)	Peru	National	Any toilet (not open defaecation) access	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Gamper-Rabindran et al. (2008)	Brazil	National	Sewage connection	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Gebre et al. (2011)	Ethiopia	Rural	Sanitation (latrine slab provision, latrine promotion)	All-cause mortality	12-59s All ages	Unimproved	Unimproved	Cluster-RCT	High risk
Geruso and Spears (2018, 2019)	India	National	Non-open defaecation Non-open defaecation among neighbours	All-cause mortality	0-11s	Improved	Unimproved	NRS	High risk
Gyimah (2002)	Ghana	National	Flush toilet provision	All-cause mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Hoque et al. (1999)	Bangladesh	Rural	Presence of faeces around latrine	Diarrhoea mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Hoque et al. (1999)	Bangladesh	Rural	Presence of faeces around latrine	Mortality - infectious disease Mortality due to ARI	0-59s	Unimproved	Unimproved	NRS	High risk
Howlader and Bhuiyan (1999)	Bangladesh	National	Flush toilet	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Luby et al. (2018)	Bangladesh	Rural	Latrine provision	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	Some concerns

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Macassa et al. (2004)	Mozambique	National	Non-open defaecation	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Masset and White (2003)	India	State-wide	Safe sanitation access	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Mellington and Cameron (1999)	Indonesia	National	Latrine access	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Null et al. (2018)	Kenya	Rural	Sanitation (latrine provision and potties)	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	High risk
Victora et al. (1988)	Brazil	Urban	Flush or pit latrine ownership	Diarrhoea mortality	0-11s	Unimproved	Unimproved	NRS	High risk
Hygiene									
Bowen et al. (2012)	Pakistan	Urban	Hygiene promotion	All-cause mortality	0-95s	Improved	Improved	Cluster-RCT	High risk
Cole et al. (2012)	South Africa	Urban	Soap, detergent and health education	All-cause mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Ercumen et al. (2015a)	Bangladesh	Rural	Safe storage	All-cause mortality	6-30s	Improved	Unimproved	RCT	Some concerns
Gyorkos et al. (2013)	Peru	Urban	Hygiene education	All-cause mortality	120s	Unimproved	Unimproved	Cluster-RCT	Some concerns
Luby et al. (2004)	Pakistan	Urban	Antibacterial soap provision	All-cause mortality	0-35s	Unimproved	Improved	Cluster-RCT	Some concerns
Luby et al. (2004)	Pakistan	Urban	Plain soap provision	All-cause mortality	0-35s	Unimproved	Improved	Cluster-RCT	Some concerns
Luby et al. (2006)	Pakistan	Urban	Soap provision	All-cause mortality	All ages	Unimproved	Improved	Cluster-RCT	Some concerns
Luby et al. (2018)	Bangladesh	Rural	Handwashing station and soap provision	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	Some concerns

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Nicholson et al. (2014)	India	Urban	Soap provision and social marketing	All-cause mortality	60-71s	Improved	Unimproved	Cluster-RCT	High risk
Null et al. (2018)	Kenya	Rural	Handwashing station and soap provision	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	High risk
Ram et al. (2017)	Bangladesh	Rural	Handwashing station and promotion	All-cause mortality	0-1s	Improved	Unimproved	RCT	Some concerns
Rhee et al. (2008)	Nepal	Rural	Handwashing with soap and water	All-cause mortality	0-1s	Improved	Unimproved	NRS	High risk
Multiple WASH									
Bowen et al. (2012)	Pakistan	Urban	Hygiene promotion and household water treatment	All-cause mortality	0-95s	Improved	Improved	Cluster-RCT	High risk
Clasen et al. (2014)	India	Rural	Sanitation promotion (CLTS), subsidies and marketing, and hygiene	All-cause mortality	0-59s All ages	Improved	Unimproved	Cluster-RCT	Some concerns
Galdo and Briceño (2005)	Ecuador	Urban	Piped water supply and sewer connection	All-cause mortality	0-59s	Improved	Unimproved	NRS	High risk
Granados and Sańchez (2013)	Colombia	National	Decentralised water supply and sewer connection	All-cause mortality	0-11s	Improved	Improved	NRS	High risk
Instituto Apoyo (2000)	Honduras	Rural	Water supply, latrines and sewer connection	All-cause mortality	0-59s	Improved	Improved	NRS	High risk
Luby et al. (2006)	Pakistan	Urban	Household water treatment (flocculant) and soap provision	All-cause mortality	All ages	Unimproved	Improved	Cluster-RCT	Some concerns
Luby et al. (2018)	Bangladesh	Rural	Household water treatment, latrines and handwashing	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	Some concerns

Study	*Country*	*Location*	*WASH intervention technology*	*Outcome*	*Age group*	*Baseline water*	*Baseline sanitation*	*Design*	*Bias in mortality estimate*
Messou et al. (1997)	Côte d'Ivoire	Rural	Source water supply and latrine provision, handwashing promotion	Diarrhoea mortality	0-59s	Unimproved	Unimproved	NRS	High risk
Null et al. (2018)	Kenya	Rural	Household water treatment, sanitation and handwashing	All-cause mortality	0-23s	Improved	Unimproved	Cluster-RCT	High risk
Pickering et al. (2015)	Mali	Rural	Sanitation promotion (CLTS) and hygiene	All-cause mortality	0-59s All ages	Unimproved	Unimproved	Cluster-RCT	Some concerns
Rasella (2003)	Brazil	Urban	Water supply and sanitation	All-cause mortality Diarrhoea mortality	0-59s	Improved	Improved	NRS	High risk
Reese et al. (2019)	India	Rural	Piped water supply, latrines and handwashing	All-cause mortality	0-59s All ages	Unimproved	Unimproved	NRS	Some concerns
Semenza et al. (1998)	Uzbekistan	Urban	Household water treatment (chlorine), safe storage and hygiene education	Diarrhoea mortality	0-59s	Unimproved	Improved	RCT	High risk

In Clasen et al. (2016), migrants were assumed to be those families who dropped out. Pickering et al. (2015) reported the total number of households who migrated, merged with other households or could not be located, which were all assumed to be migrants.[152] In Crump et al. (2005), information was given on person-weeks "missing because of short or long term migration" (p.2), while Bowen et al. (2012) and Null et al. (2018) reported numbers lost or absent, which were assumed to be permanent migrants.

Age-specific mortality rates for children were calculated by replacing equation (5.3) with the numbers of deaths and population shares among children. Cause-specific mortality rates were calculated by replacing D_j with numbers of deaths attributed to diarrhoea and/or infectious diseases, determined by recalled verbal autopsy or taken from vital registration data. Vital registration and verbal autopsy estimates are also used in GBD calculations (GBD Cause of Death Collaborators, 2017c). An important issue affecting crude death rate calculations is that they are right-censored; that is, where data are collected contemporaneously among participants regardless of age, children born into the study and younger children have completed shorter durations than older children. This causes downwards bias in the estimate of mortality in any single trial arm, although the bias may be less problematic in randomised trials with contemporaneous data collection across arms. In the case of Null et al. (2018), households were eligible for inclusion in the study where women reported being pregnant (in the second or third trimester) during the pre-allocation census, and outcomes were collected on children at age 2.[153] Hence, in this case, the under-2 and neonatal mortality rate (MR) per 1,000 live births was calculable, which is not susceptible to censoring:

$$MR_j = \frac{D_j}{\left(B_j - B_j^D\right)} \; x \; 1{,}000 \qquad (6.4)$$

where B_j is the number of live births and B_j^D the number of still-births. In practice, age-specific crude death rates and U2MR estimates were almost

[152] These figures needed to be adjusted, respectively, by the reported share of children and average number of household members, in order to estimate total numbers of child migrants and total population in Pickering et al. (2015).
[153] Luby et al. (2018) also recruited participants in the first two trimesters, measuring outcomes at median follow-up of 22 months, with inter-quartile range 21-24 months.

identical and the results were unaffected when using either estimate. In one case, Ram et al. (2017) which followed up neonates for one month, the hygiene intervention commenced during the prenatal period therefore the crude mortality rate calculation was used including still-births and neonatal deaths.

Where studies reported independent treatment and control arms, data for mortality from each treatment-control comparison were included. However, many studies reported multiple correlated effect sizes. For example, factorial studies compared multiple treatment groups against a single control (e.g., Luby et al., 2018; Null et al., 2018). Others reported data separately for multiple age groups (e.g., Gebre et al., 2011). There are two fundamental problems in including multiple effect estimates from any one study in a single meta-analysis (Higgins et al., 2011). First, studies with multiple results would receive greater weight than studies with only one effect estimate. Second, the effect estimates from multiple treatment arms with a single control group are positively correlated, and not accounting for this positive correlation leads to the underestimation of summary variance (Borenstein et al., 2009). Where studies reported multiple treatments compared to a single control arm, so the comparisons were not independent, the control arms were split by assuming the populations and deaths were evenly distributed between comparisons (affecting the precision of estimate, but not the effect size).

6.5 Critical appraisal

Comprehensive critical appraisal was done, including risk-of-bias and publication bias assessment. Drawing on the tool presented in this Thesis (developed in Chapter 4 and presented in Appendix A), risk of bias was assessed according to confounding, selection bias, deviations from intended intervention (including performance bias and measurement of intervention), attrition bias, outcome measurement error, and bias in reporting results. No studies were found to have 'low risk of bias' in attributing changes in mortality to the intervention. It is important to emphasise, however, that the studies were critically appraised on the likelihood of bias in estimating effects of WASH access on mortality, which may or may not have been a primary research question in the papers themselves. For example, Geruso and Spears (2018, 2019) presented a natural experiment in India where the main question of interest was in explaining the perverse relationship between

socioeconomic status and infant mortality between Hindu and Muslim communities, and not in estimating the effect on mortality of latrine access or water supply. Risk-of-bias assessments are reported separately for RCTs (Figure 6.3) and NRS (Figure 6.4). The full appraisals by study are reported in Appendix D (Tables D1 and D2).

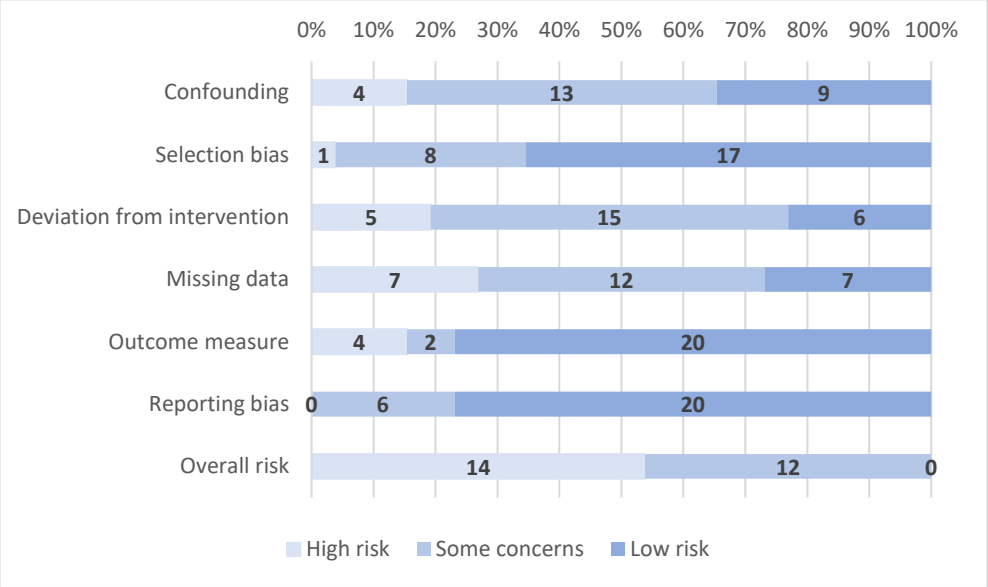
6.5.1 *Risk of bias for RCTs*

One-third of RCTs reported using adequate allocation sequence and concealment, and demonstrated pre-test covariate balance, to satisfy a 'low risk' rating on confounding. A study which assigned households alternately by field workers (Conroy et al., 1999) and did not present covariate balance was rated of 'high risk of bias'. Others used adequate randomisation but presented pre-test covariate imbalances that were beyond those expected by chance, or imbalances in access to water supply (Bowen et al., 2012; Emerson et al., 2004; Gebre et al., 2011; Nicholson et al., 2014), quality of water (Crump et al., 2005), water treatment and storage practices (Boisson et al., 2010; Jain et al., 2010; Lule et al., 2005), sanitation (Boisson et al., 2010; Crump et al., 2005; Pickering et al., 2015) or hygiene practices (Lule et al., 2005). Imbalances appeared to be related to sample size in some cases. For example, Gebre et al. (2011) randomised 24 clusters, 12 to receipt of latrine slabs and latrine construction training, and 12 to control. In some cases, data were collected on water, sanitation and hygiene at pre-test, but balance was not presented for all variables (Morris et al., 2018, for sanitation and hygiene; Ram et al., 2017, for sanitation). Crump et al. (2005) did not indicate how randomisation was done, but due to the involvement of Stephen Luby, who had by that time already published studies where the randomisation process was clearly described, the study was assumed to have used adequate sequence generation and allocation concealment. These were all assessed as having 'some concerns' due to confounding.

Risk of selection bias related to the timing of individual participant recruitment with respect to treatment allocation. Where participants were recruited before allocation in cluster-RCTs, or where recruiters were blinded to allocation, the studies were judged to be of 'low risk of bias'. Where recruitment was done afterwards by those potentially with knowledge of allocation (e.g., Luby et al., 2004, 2006; Nicholson et al., 2014), or where

individuals needed to be recruited later due to attrition (losses to follow-up during the trial) (e.g., Clasen et al., 2014; Pickering et al., 2015).

Figure 6.3 Overall risk-of-bias assessments for included RCTs



Deviations from intended interventions were due to factors relating to motivation bias, such as where data were collected weekly (Luby et al., 2004, 2006; Lule et al., 2005) or bi-weekly (Nicholson et al., 2014) over the course of a year or more, possible contamination or substitution effects among controls (Lule et al., 2005), or the apparent effectiveness of placebo interventions (Boisson et al., 2010; Jain et al., 2010). In the case of Boisson et al. (2010), who provided controls a placebo LifeStraw water filter with tap, with instructions that participants "drink filtered water directly from the tap and not to store filtered water in order to prevent recontamination" (p.3), an explanation for the apparent effectiveness of the placebo was the safe storage inherent in the device. Perhaps this may also help explain why filtration has been found to be more effective than other water treatment approaches in reducing diarrhoeal morbidity in recent systematic reviews and meta-analyses (e.g., Hunter, 2009; Clasen et al., 2015; Wolf et al., 2018).

In general, contamination or spillover effects were judged unlikely to be problematic where studies used cluster-randomisation or reported geographical separation of participants (e.g., Emerson et al., 2004; Gebre et al., 2011; Luby et al., 2018). Of specific relevance to mortality estimates, several studies provided ORS to severely ill children and/or encouraged

mothers to attend health clinic (Ercumen et al., 2015a; Jain et al., 2010; Luby et al., 2004, 2006; Lule et al., 2005; Mengistie et al., 2013; Peletz et al., 2012).

Studies with attrition rates greater than 20 percent, with no information provided about reasons for drop-outs by intervention group, tests for covariate balance or robustness of findings, were assessed as being of 'high risk of bias' (Bowen et al., 2012; Conroy et al., 1999; Du Preez et al., 2011). One study also reported 10 percentage points higher attrition in control group than treatment (Gebre et al., 2011).

Cause-specific mortality determined by participant verbal autopsy is more likely to be biased than all-cause mortality (Wood et al., 2008; Savović et al., 2012).[154] All-cause mortality was usually categorised as being a reliable outcome even if it was self-reported, providing the recall period was under one month. If cause-specific mortality was measured, assessment was made as to whether it was self-reported (Pickering et al., 2015). In addition, one study collecting reported all-cause mortality used a six-year recall (Bowen et al., 2012).

A striking finding from the trials is that only one reported finalising a pre-analysis plan (Peletz et al., 2012), and only two reported blinding data analysts to intervention (Luby et al., 2018; Null et al., 2018). 'Some concerns' were raised about selective reporting in Gebre et al. (2011), where cause-specific mortality was not reported by intervention groups despite verbal autopsies being taken, or where studies did not report having a trial registry (Conroy et al., 1999; Emerson et al., 2005; Lule et al., 2005).

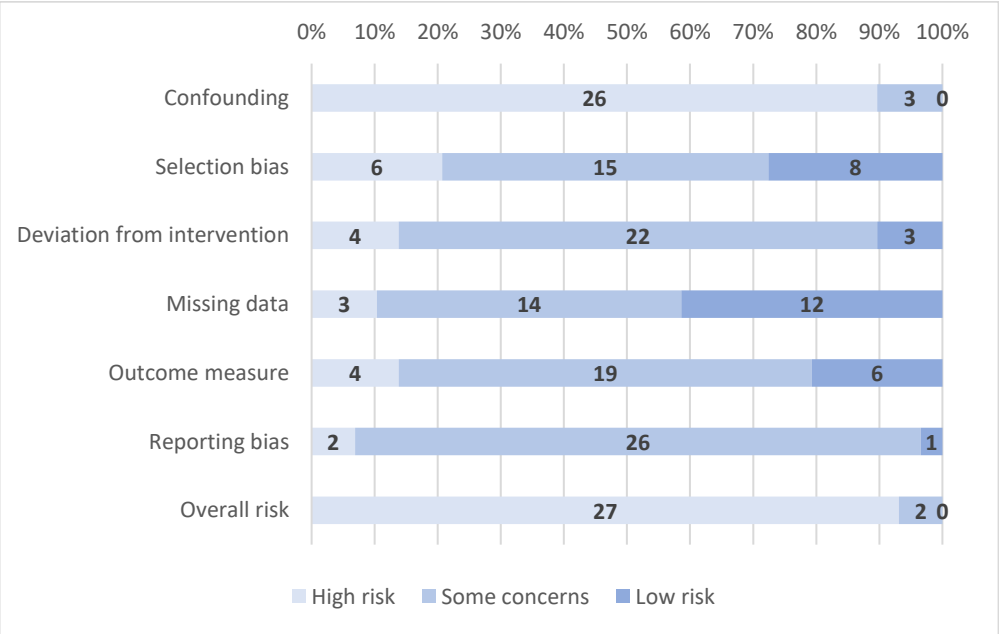
6.5.2 *Risk of bias for NRS*

Only three NRS were assessed as having 'some concerns' about confounding, studies of privatised water provision in Argentinean municipalities (Galiani et al., 2005), improved water supply reliability in India (Ercumen et al., 2015b), and piped water supply and latrines in India (Reese et al., 2019) (Figure 6.4). In all cases, participation was deemed largely determined by

[154] However, the study that included a passive control, receiving no between survey follow-up visits by health promoters, and an active control, receiving follow-up visits but no WASH hardware intervention, found an odds ratio of all-cause mortality among under-2s of 0.88 favouring the active control group (95%CI=0.58, 1.37), suggesting there may have been differences in reporting incentives for all-cause mortality had the study been powered to detect it with statistical precision.

programme placement – in Galiani et al. (2005) the local government's decision to privatise, in Reese et al. (2019) all households in a community were simultaneously connected to the water supply by the NGO Gram Vikas, while in Ercumen et al. (2015b) all households were connected to the municipal supply. Participation was then carefully modelled using a rich set of covariates measured at baseline and based on theory and factors influencing programme targeting (e.g., whether the municipality was led by the ruling party implementing the reforms in Galiani et al., 2005). The estimations ensured common support through statistical matching. Galiani et al. (2005) incorporated baseline outcome measurement in double differences, which they supported by presenting common and equal trends for five years prior to reforms being implemented, which was formally tested using a leads and lags model. Both studies presented null results for placebo outcomes, mortality due to non-infectious causes in Galiani et al. (2005), and prevalence of bruising and scrapes in Ercumen et al. (2015b) and Reese et al. (2019). They also showed how the reforms had led to improvements in WASH access using causal pathway analysis. Ercumen et al. (2015b) and Reese et al. (2019) also reported various health outcomes, such as diarrhoea carer-report, helminth infection from stool samples and anthropometry (Reese et al., 2019) and bloody diarrhoea and typhoid (Ercumen et al., 2015b).

Figure 6.4 Overall risk-of-bias assessments for included NRS



All other NRS were judged to have 'high risk of bias' due to confounding, where participation was not modelled using pre-test covariates or a rich set of covariates based on knowledge of programme allocation decisions. This was the case for all studies based on demographic and health survey (DHS) data (e.g., Brockerhoff, 1990; Brockerhoff and Derose, 1996; Casterline et al., 1989; DaVanzo and Habicht 1986; Fink et al., 2011; Howlader and Bhuiyan 1999; Mellington and Cameron, 1999). In the case of prospective NRS, pre-test covariate imbalance was either beyond that expected due to chance (Cole et al., 2012), or was not given (Rasella, 2003).

Selection bias and attrition bias were deemed less problematic where studies used census data (Galdo and Briceño, 2005; Galiani et al., 2005; Gamper-Rabindran et al., 2008), census-based random sample survey data, like DHS (Abou-Ali et al., 2010; Fink et al., 2011), or vital registration (Granados and Sańchez, 2013; Rasella, 2003; Victora et al., 1988). There were concerns about selection bias when a prospective study recruited openly (Cole et al., 2012). One retrospective study used migration status as an identification variable in the participation equation, which may have been endogenous since, according to the paper, property values were observed by one long-term resident to increase due to the water and sanitation improvements made under the project (Galdo and Briceño, 2005). It appeared that some cohort studies included children born into the study during analysis, which may lead to selection bias due to right censoring (Ercumen et al., 2015b; Reese et al., 2019). A few DHS studies were able to address this source of selection bias through proportional hazards regression (Brockerhoff, 1990; Brockerhoff and DeRose, 1996; Gyimah, 2002; Masset and White, 2003) or by restricting observations to infants born more than 12 months prior to the survey (Howlader and Bhuiyan, 1999).

Concerns about deviations from intended interventions usually related to measurement of the technology received in retrospective studies, for example where WASH was measured as a self-reported exposure (e.g., Fuentes et al., 2006; Fink et al., 2011; Gamper-Rabindran et al., 2008; Geruso and Spears, 2018; Rasella, 2003; Rhee et al., 2008) and therefore susceptible to over-reporting (Briscoe et al., 1985). Prospective studies examining child mortality are limited due to ethical reasons required to measure it accurately, such as the need to withhold curative treatment such as oral rehydration. One study did provide ORS in treatment areas only, causing likely overestimation

of differences in mortality due to WASH (Messou et al., 1997). The two case-control studies were assessed as being of 'low risk of bias', where the authors conducted spot checks to confirm reported access (Hoque et al., 1999; Victora et al., 1988). In addition, data collectors were blinded to the 'assignment' in Victora et al. (1988) since cause of death was only collected after observing WASH access.

As with RCTs, concerns about mortality measurement usually related to the length of recall in survey or census data. In the case of Victora et al. (1988), a monitoring system was set up to collect all infant mortality data in the city over a 12-month period, including weekly visits to hospitals, coroners and death registries. For each infant death due to infectious disease, or death of unknown cause, a physician visited the family to collect information about the terminal illness.

Finally, the issue with all retrospectively designed NRS is that authors may be more liable to decisions about analysis and reporting based on findings. Only one study (Reese et al., 2019) was of 'low risk of bias', because it pre-registered and published a baseline report with pre-analysis plan (Reese et al., 2017). Another provided a protocol as a supplementary file to the published study but indicated that the decision had been taken to collect mortality for under-2s afterward the protocol was filed (Ercumen et al., 2015b). Two studies were deemed to have probably determined WASH technology variables based on findings (Hoque et al., 1999, for water storage cut-off at 2 litres; Fuentes et al., 2006).

6.5.3 *Analysis of publication bias*

Analysis was undertaken of publication bias using standard approaches (Egger et al., 1997; Peters et al., 2008). Publication bias occurs if the outcome of the study affects the likelihood (or speed) of publication, with the result that findings in the published literature are systematically unrepresentative of the population of studies (Rothstein et al., 2005). The bias is usually in favour of positive (reductions in mortality) and significant findings (Dickersin, 1990). Publication bias is related to outcome reporting bias, assessed under risk-of-bias, where only those outcomes supporting the researchers' priors are reporting, but it also incorporates biased exploratory analysis (p-hacking) and full suppression of findings (file-drawer effects).

Publication bias is thought to be a potential source of bias in diarrhoea morbidity studies (e.g., Curtis and Cairncross, 2003; Waddington et al., 2009). It is especially important to analyse for the studies reviewed here, warranted for the non-randomised studies for the usual reasons that studies which aim to estimate a relationship primarily on child mortality are more likely to be published if they find a significant effect. Many of these studies were produced by demographers and econometricians. Suspicion about the representativeness of findings published in economics journals goes back at least to the 1970s (Leamer, 1978). Small-sample studies that show low statistical significance are at a disadvantage to publication selection in empirical economics research (Stanley, 2005) and health (Easterbrook et al., 1991; Vickers et al., 1998; Hopewell et al., 2009). The exploratory social science research tradition and, until recently, limited production of study protocols or pre-analysis plans suggests there are potentially severe problems of publication bias due to 'p-hacking' to find statistically significant findings, and this problem may arise particularly in studies of observational data. But publication bias due to file-drawer effects may also be partly mitigated by the traditions of publication in development research – for example working papers in economics and political science – and modern electronic dissemination (Duvendack et al., 2012; Rothstein et al., 2005).

No RCTs were adequately powered to analyse mortality outcomes with statistical precision. Since the mortality data were collected from participant flow diagrams, the fact that mortality estimates are available at all is indicative of the improved quality of reporting in these studies, following best-practice guidance (Moher et al., 1998). This suggests publication bias may be limited for the prospective studies including the RCTs. However, it is also possible that reporting of mortality is censored in the sample of RCTs contained here – since those studies where zero children died over the course of the study, which would have contributed an equivalent odds ratio of 1, if that were calculable, or a risk difference equal to zero, were omitted.

Publication bias analysis was done using two methods. Direct tests for publication bias were done in meta-regression accounting for whether the study was published in a peer-review journal. Indirect testing of small-study effects used graphical inspection of funnel plots (Peters et al., 2008) and formal regression tests (Egger et al., 1997). These tests assume that there are

weaker incentives for researchers and journals to publish smaller sample studies that do not show significant findings, because the cost of such studies is less and/or that authors of underpowered (small-sample) studies are more likely to undertake p-hacking in order to obtain publishable results.

The results of publication bias analysis (Table 6.4) suggested small-study effects were evident for NRS, both for all-cause mortality and diarrhoea mortality. They were also evident for exposure studies for all-cause mortality, but not for intervention studies. In Figure 6.5, the contour-enhanced funnel graphs (Peters et al., 2008) for RCTs and NRS are overlain with the Egger's test regression lines, indicating clear asymmetry and clustering of NRS in areas of statistical significance, together with a negative intercept coefficient on the regression of effect size on standard error (funnel graph axes are inverted). This is consistent with publication bias due to small-study effects.

Table 6.4 Publication bias assessment

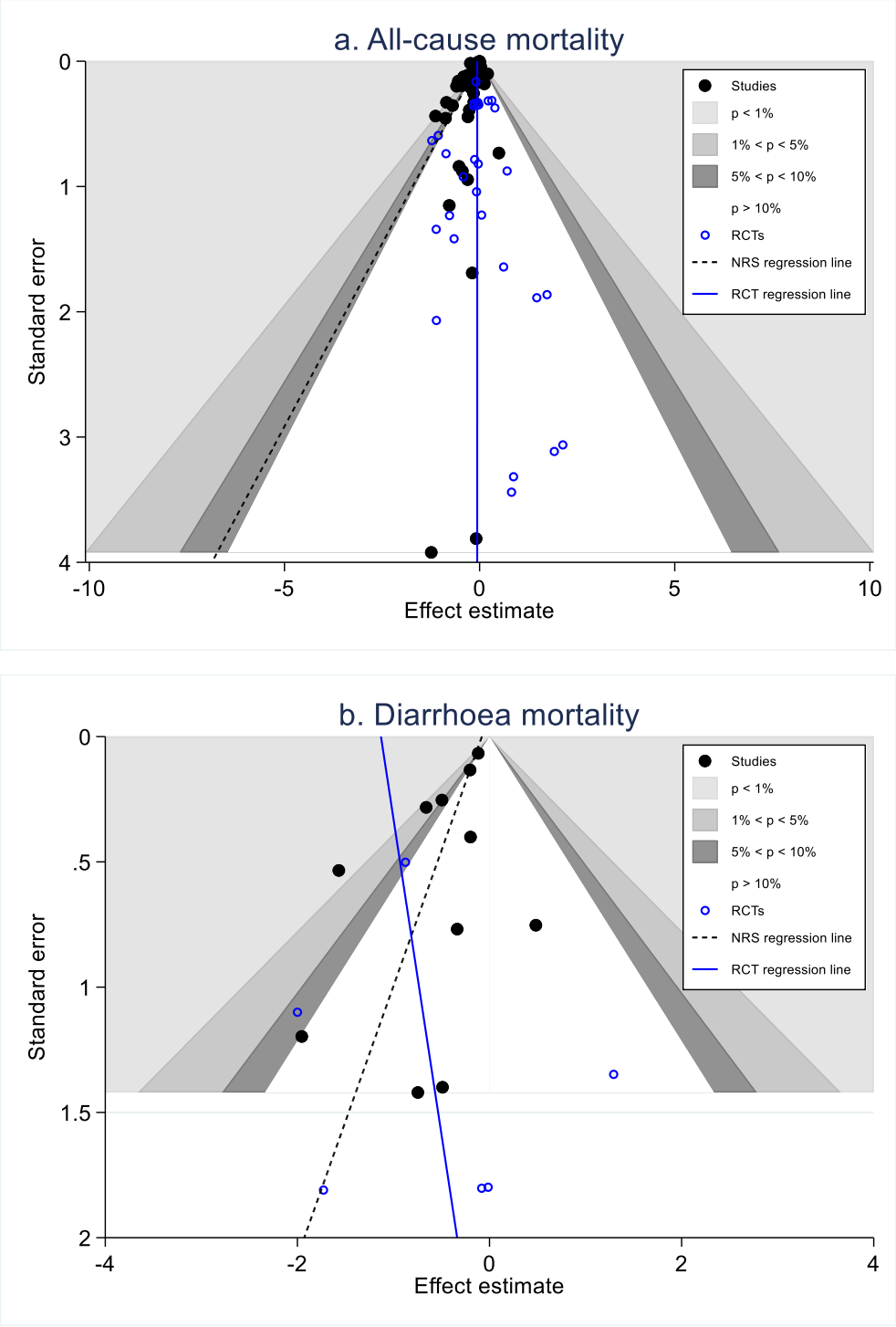
Outcome	*Analysis*	*Test*	*Coefficient*	*t*	*# obs*
All-cause mortality	All studies	Egger	0.36***	-3.88	79
	RCTs	Egger	0.94	-0.32	31
	NRS	Egger	0.18***	-4.15	47
	Intervention study	Egger	1.01	-0.39	38
	Exposure study	Egger	0.15***	-4.15	40
	Publication bias^	1=journal article	0.99	-0.32	79
Diarrhoea mortality	All studies	Egger	0.49	-1.94	17
	RCTs	Egger	0.40	0.49	6
	NRS	Egger	-0.92*	-2.20	11
	Intervention study	Egger	0.77	-0.39	8
	Exposure study	Egger	0.48	-1.48	9

Notes: Egger test reports exponentiated intercept coefficients e^b; ^ meta-regression of all-cause mortality on publication status (no studies of diarrhoea mortality were published outside of peer-review journals); *** p<0.001, * p<0.1.

In contrast, RCTs are generally symmetrically distributed, as confirmed by the regression line indicating near-zero intercept coefficient on the regression of the effect on study size for all-cause mortality (Table 6.4). The regression line for RCTs reporting diarrhoea mortality suggests a reverse small-study effect – that is, RCTs were more likely to report deaths when they were less powered to do so. Evidence does not suggest significant small-study

effects for intervention studies either.[155] In sum, there is strong evidence for publication bias in NRS but not RCTs. This may be expected since none of the RCTs were designed to estimate the effects on mortality, whereas all NRS estimated effects on mortality as a primary outcome.

Figure 6.5 Funnel graphs for mortality with regression lines



[155] Funnel graphs for intervention studies are given in Appendix D Figure D1.

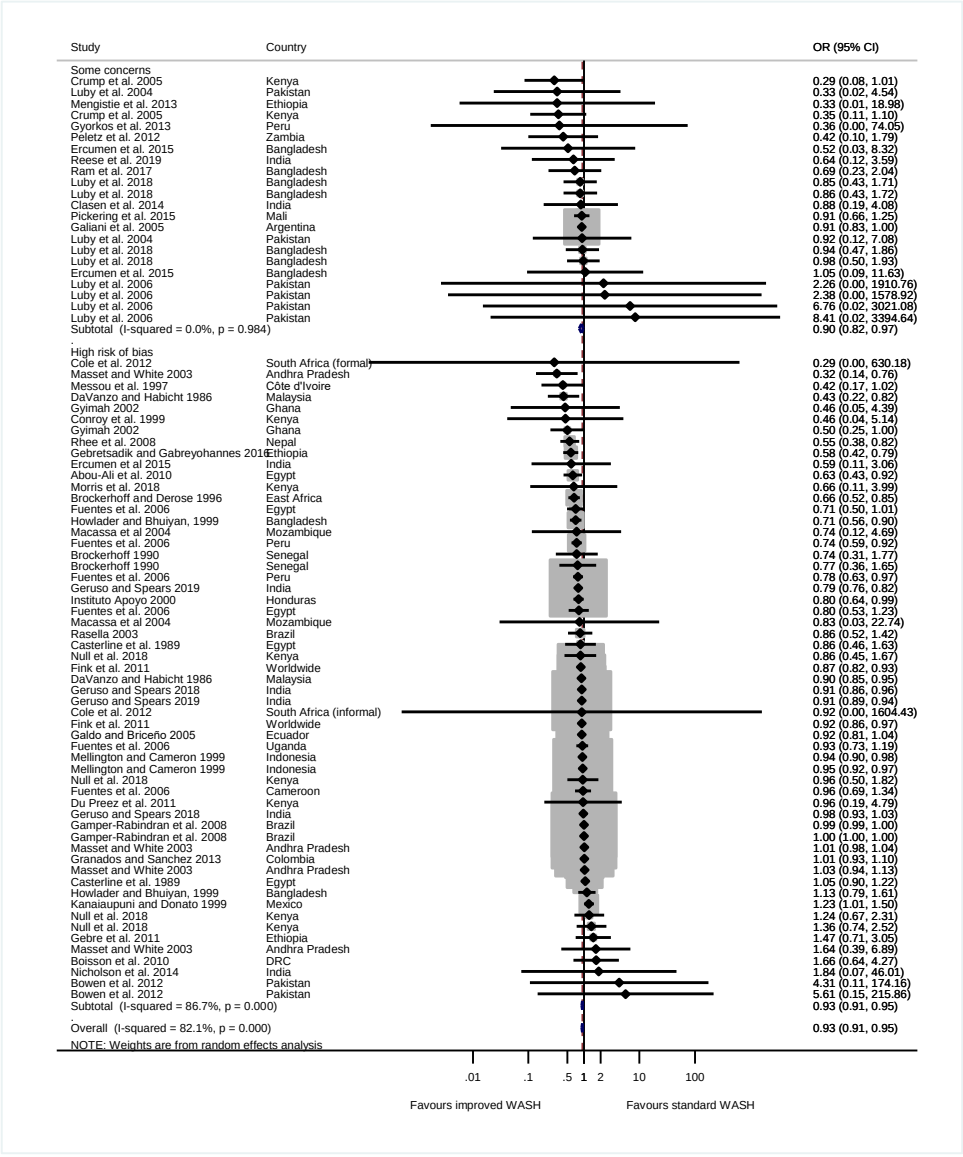
6.6 Meta-analysis results

Random effects meta-analysis was used to synthesise the findings. A standard approach to meta-analysis was followed, including sensitivity analysis by risk-of-bias status, sub-group analysis by mortality causation, bivariate moderator analysis and multivariate meta-regression. Moderator variables were pre-specified, based on what might theoretically be associated with mortality, and a general-to-specific approach was taken to determine the optimum meta-regression specification (Mukherjee et al., 1997).

The particular value of meta-analysing mortality outcomes across studies is the increased power provided from the synthesis of multiple findings, enabling statistical precision even for small effect sizes (Greenhalgh, 2014). Out of 77 effects on all-cause mortality included in analysis, 21 were individually significant at 95 percent confidence, all of which were from NRS; none of the 31 RCT arms was powered to estimate an individually significant effect on all-cause mortality. The results of bivariate meta-analysis of all-cause mortality suggest a significant reduction in the odds of death (OR=0.93; 95%CI=0.91, 0.95) (Figure 6.6), measured across 77 study arms, with a p-value 'to die for' of $p<10^{-10}$. Although the unexplained proportion of the variance across studies is relatively high (I-squared=82%), the estimated magnitude of statistical heterogeneity is low (tau-squared=0.0009, or odds fewer than 1 in 1,000). Nevertheless, further analyses were undertaken to examine the sensitivity of the findings and attempt to explain the residual between-study variation.

Sensitivity analysis was done, firstly, according to risk-of-bias rating. There was 10 percent reduction in odds of child mortality associated with improved WASH access for the 22 study arms with 'some concerns', with 95 percent confidence interval between 3 and 18 percent (OR=0.90, 95%CI=0.82, 0.97) (Figure 6.6). The difference in odds ratios for pooled effects of studies with 'some concerns' and those with 'high risk of bias' was not significant ($p<0.35$). The unexplained component of the variance in the estimate for studies with 'some concerns' is zero (I-squared=0%). The statistical heterogeneity in findings for studies with 'high risk of bias' was also low (I-squared suggests high between study variation at 87%, but tau-squared indicates the magnitude of that variation is small at 0.0009).

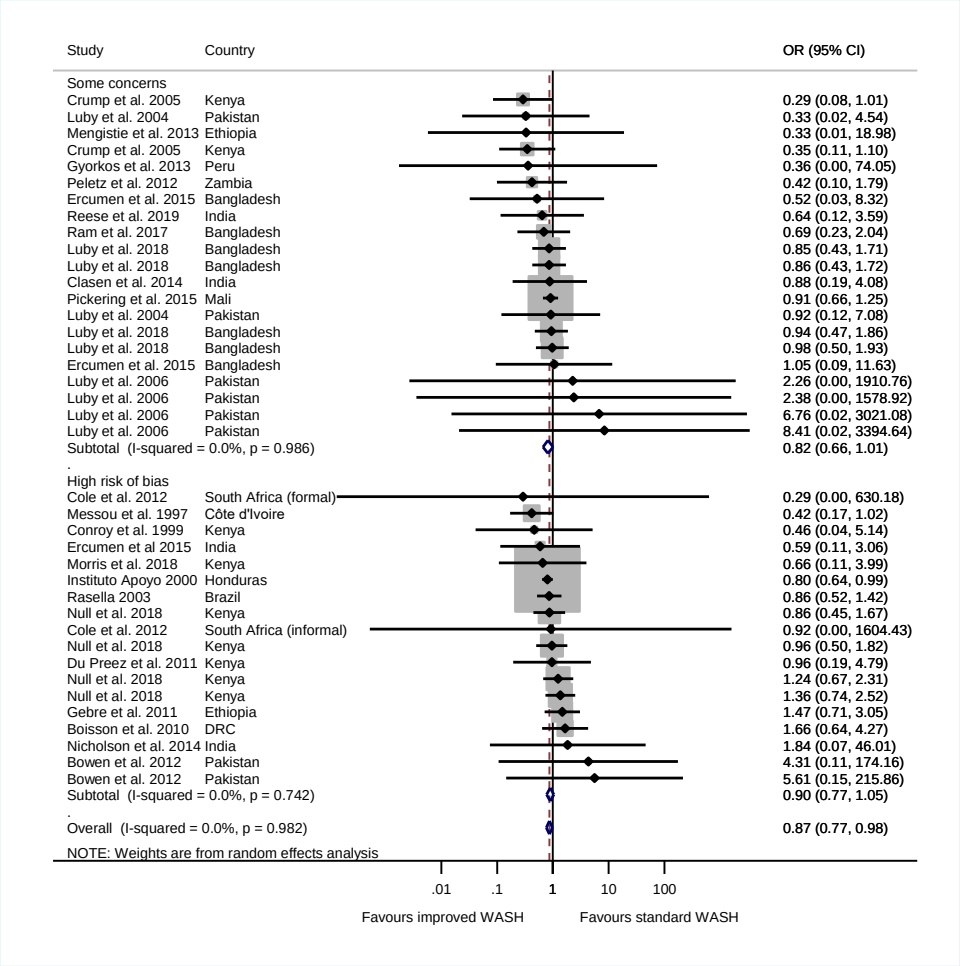
Figure 6.6 All-cause mortality in childhood



The sensitivity of the findings to study design and other factors that may be associated with the estimated effect was also explored. Restricting the analysis to intervention studies, the findings suggested a 13 percent reduction in all-cause mortality in childhood on average (OR=0.87, 95%CI=0.77, 0.98; I-squared=0%; tau-squared=0.000) (Figure 6.7). For RCTs, studies with 'some concerns' were associated with 18 percent reduction in all-cause mortality due to improved WASH (OR=0.82, 95%CI=0.67, 1.02; I-squared=0%; tau-squared=0.000) (Appendix D Figure D2). This is consistent with the finding in Chapter 5, Section 5.2.2 where RCTs with 'high risk of bias' were found to estimate smaller effects than other RCTs. In addition, studies where researchers provided participants with ORS and health clinic referrals found significantly bigger reductions in childhood

mortality on average (p<0.01). Studies carried out in the rainy season also found bigger effects than those carried out at other times of year or year-round (Table 6.5).

Figure 6.7 All-cause mortality for intervention studies (RCTs and NRS)



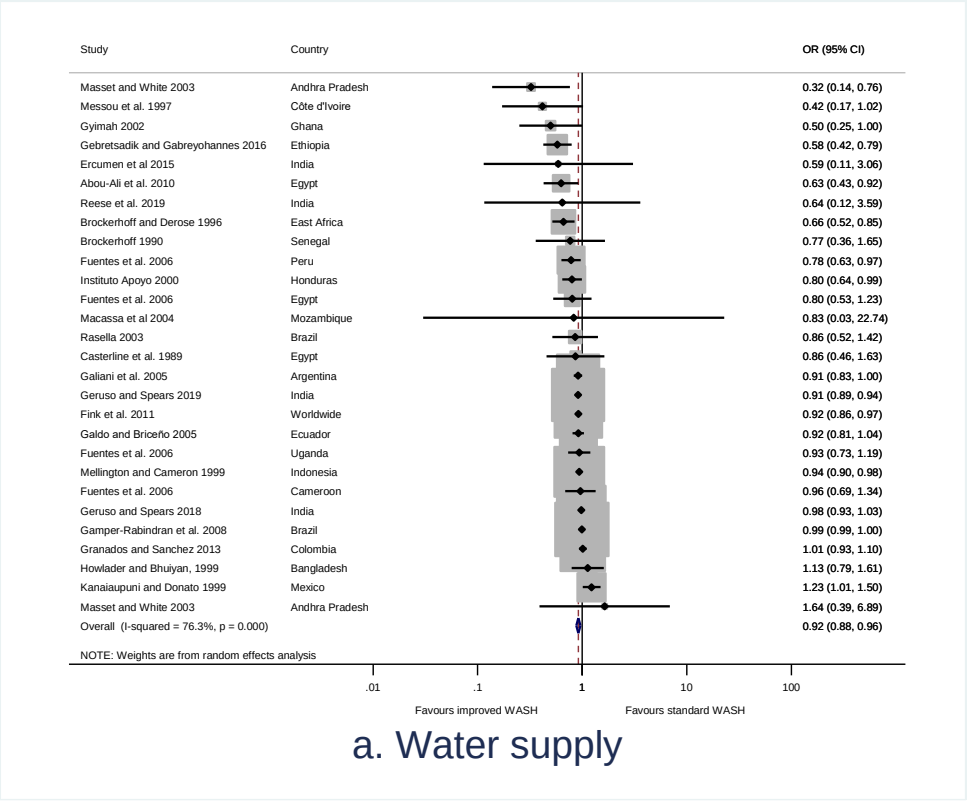
Moderator analysis was also done to examine heterogeneity in findings by WASH technology provided. Owing to the limited effectiveness of multiple WASH technologies found in previous reviews, and following Clasen et al. (2010), analysis was made of trial arms incorporating any single technology (whether done alone or alongside any other WASH technology) (Table 6.5). However, as noted above, since one reason why multiple interventions are not observed to be more effective than single interventions in diarrhoea morbidity studies is reporting bias – a factor that is not expected to be problematic for all-cause mortality – moderator analysis was also made of the main WASH intervention reported in the study (forest plot reported in Appendix D Figure D3).

Table 6.5 Sensitivity and moderator analyses: all-cause mortality

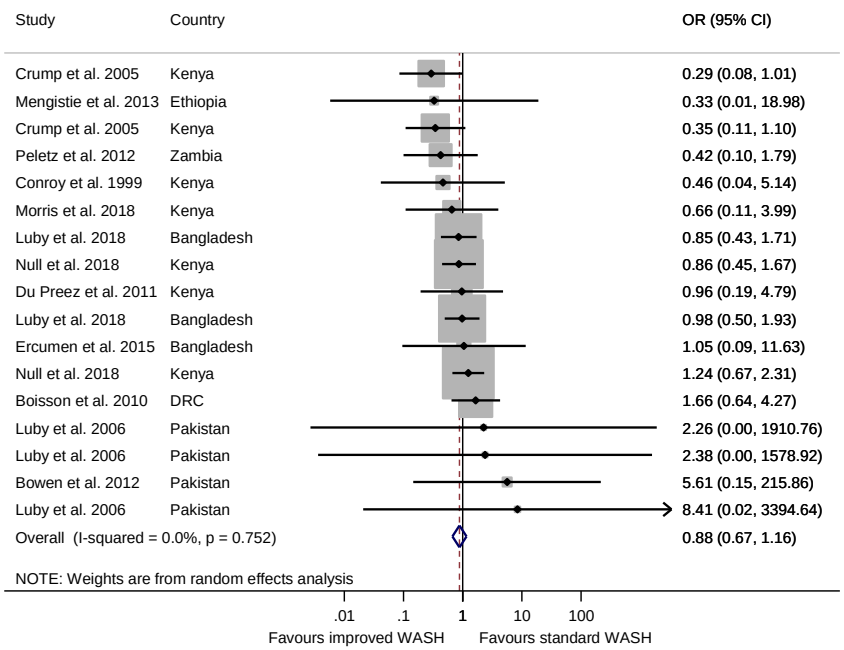
Moderator	*OR*	*95% CI*		I^2	Tau^2	*P>|z|*	*# obs*
All estimates	0.93	0.91	0.95	82%	0.001	0.000	79
Some concerns	0.90	0.82	0.97	0%	0.000	0.984	22
High risk of bias	0.93	0.92	0.95	87%	0.001	0.000	57
RCT	0.94	0.80	1.11	0%	0.000	0.975	32
NRS	0.93	0.91	0.95	89%	0.001	0.000	47
Intervention study	0.87	0.77	0.98	0%	0.000	0.982	39
Exposure study	0.93	0.92	0.95	91%	0.001	0.000	40
Rainy season	0.32	0.14	0.73	0%	0.000	0.982	3
Dry season	0.43	0.01	18.93	0%	0.000	0.973	3
Year-round	0.93	0.92	0.95	83%	0.001	0.000	73
ORS and/or health care referral	0.45	0.25	0.82	0%	0.000	0.974	12
No ORS or referral	0.93	0.92	0.95	85%	0.001	0.000	67
Water supply	0.91	0.87	0.96	80%	0.004	0.000	23
Water treatment	0.75	0.54	1.06	0%	0.000	0.764	13
Sanitation	0.91	0.86	0.97	93%	0.008	0.000	20
Hygiene promotion	0.76	0.57	1.00	0%	0.000	0.732	13
Multiple WASH	0.94	0.85	1.03	17%	0.004	0.287	10
Any water supply	0.92	0.88	0.96	76%	0.004	0.000	28
Any water treatment	0.88	0.67	1.16	0%	0.000	0.752	17
Any sanitation	0.91	0.87	0.96	91%	0.007	0.000	28
Any group sanitation	0.83	0.62	1.11	0%	0.000	0.446	4
Any hygiene promotion	0.86	0.72	1.02	0%	0.000	0.840	26
Any hygiene with improved water	0.72	0.56	0.93	0%	0.000	0.975	13
Any hygiene with unimproved water	0.98	0.78	1.24	0%	0.000	0.562	13
Baseline water improved	1.00	0.99	1.00	27%	0.000	0.059	41
Baseline water unimproved	0.90	0.86	0.94	82%	0.007	0.000	38
Baseline sanitation improved	0.95	0.90	1.01	0%	0.000	0.941	19
Baseline sanitation unimproved	0.93	0.91	0.95	86%	0.001	0.000	60

Forest plots for WASH technologies (Figure 6.8) indicate 8 percent reduction in odds of all-cause mortality associated with improved water supply on average (p<0.001; evidence from 28 studies), and 9 percent associated with sanitation (p<0.001; 28 studies). There were likely too few studies of sanitation provided to whole communities to detect significant findings (p<0.2; 4 studies). When hygiene improvements were made, all-cause mortality was reduced by 14 percent (p<0.07; 26 studies); when they were made in contexts when water supply was classed as improved according to JMP definitions, mortality was reduced by 27 percent (p<0.001; 13 studies). There was an estimated 12 percent reduction in mortality for household water treatment, but the findings were not significant (p<0.365; 17 studies). When the samples were restricted to intervention studies, the findings remained significant for water supply and sanitation, but not other technologies (Appendix D Figure D4).

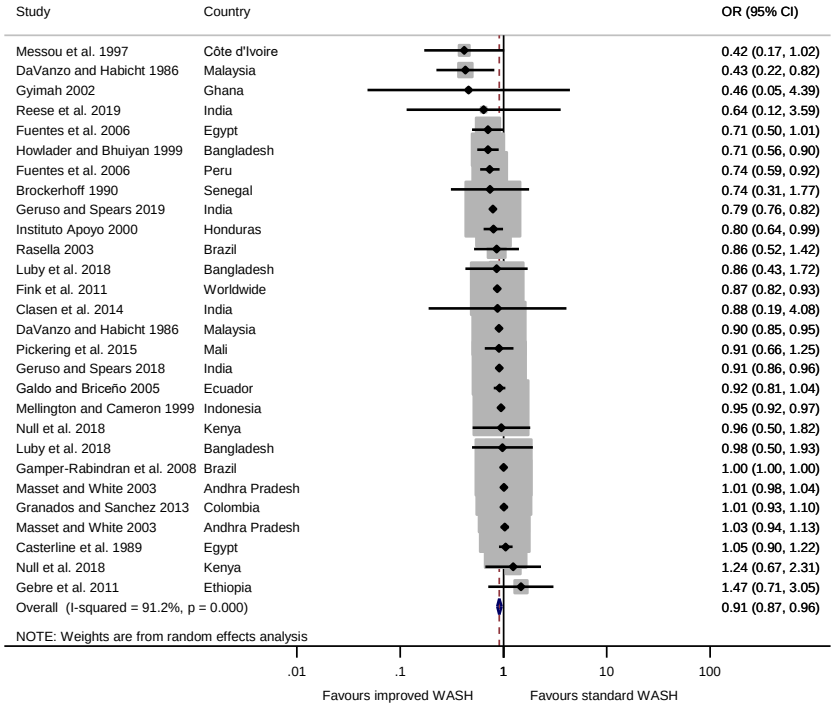
Figure 6.8 All-cause mortality by WASH technology



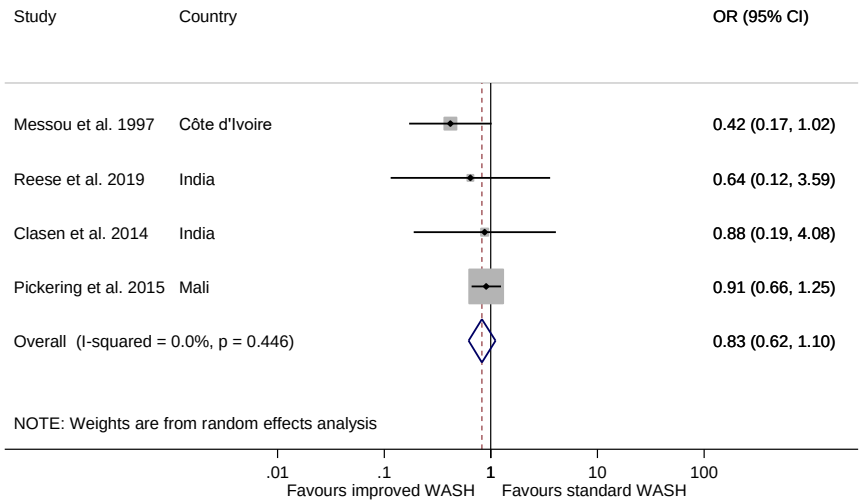
a. Water supply



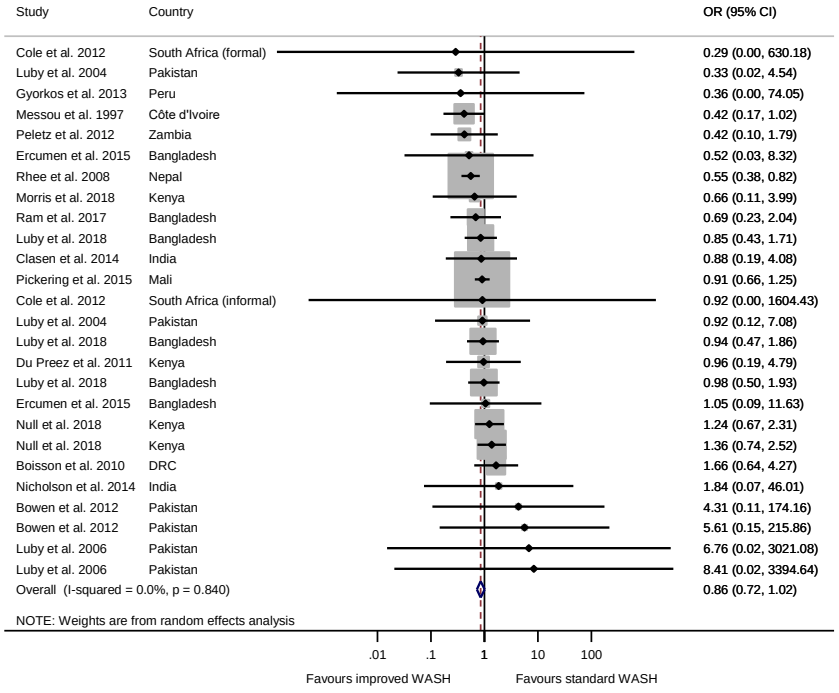
b. Water treatment



c. Household latrines



d. Latrines provided to entire community

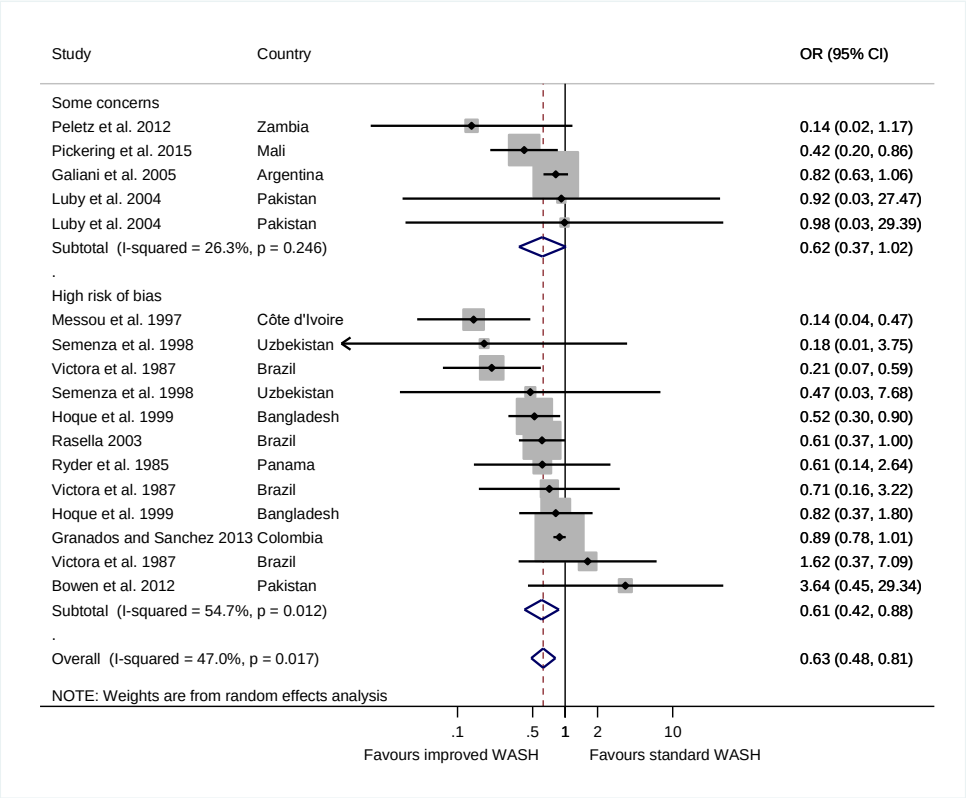


e. Hygiene

Two further analyses were performed. Firstly, analysis was done of childhood diarrhoea mortality, as determined through vital registration and verbal autopsy (Figure 6.9). Meta-analyses of all 17 studies reporting mortality by intervention group suggested WASH provision may lead to a 37 percent reduction in the odds of child death over the control mortality rate (OR=0.63, 95%CI=0.48, 0.81), or a 38 percent reduction across studies with only 'some concerns' about bias (OR=0.62, 95%CI=0.37, 1.02; p<0.06). Statistically

significant reductions in mortality were also estimated for intervention and exposure studies (Appendix D Figure D5), and RCTs with only 'some concerns' about bias (Appendix D Figure D6). Hoque et al. (1999) also reported significant reductions in mortality due to ARIs associated with water supply provision, and reductions due to other infectious diseases following water supply and sanitation provision (Appendix D Figure D7).

Figure 6.9 Childhood diarrhoea mortality

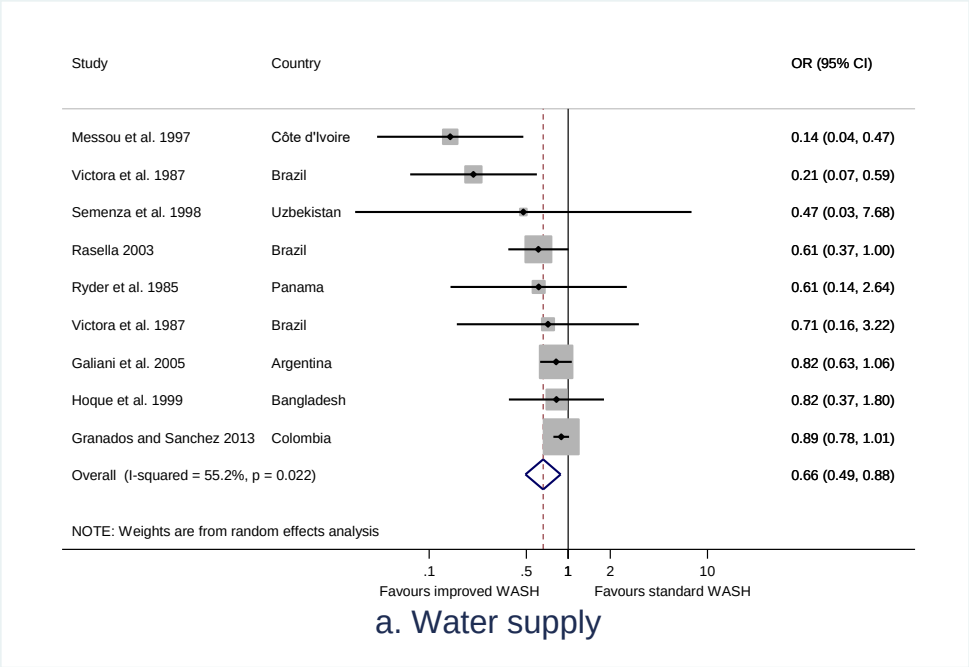


The degree of statistical heterogeneity overall suggested further exploratory analysis would be useful. In the first instance, pooled estimates were made of WASH technologies (Figure 6.10). The results indicated improved water supply (OR=0.66, 95%CI=0.49, 0.88; 9 studies), household latrines (OR=0.58, 95%CI=0.37, 0.90; 6 studies) and, especially, latrines provided to whole communities (OR=0.27, 95%CI=0.10, 0.76; 2 studies) and hygiene (OR=0.31, 95%CI=0.18, 0.55; 6 studies) caused significant and substantial reductions in childhood diarrhoea mortality.[156] Among the few studies which have been done in endemic circumstances, there was no significant effect of

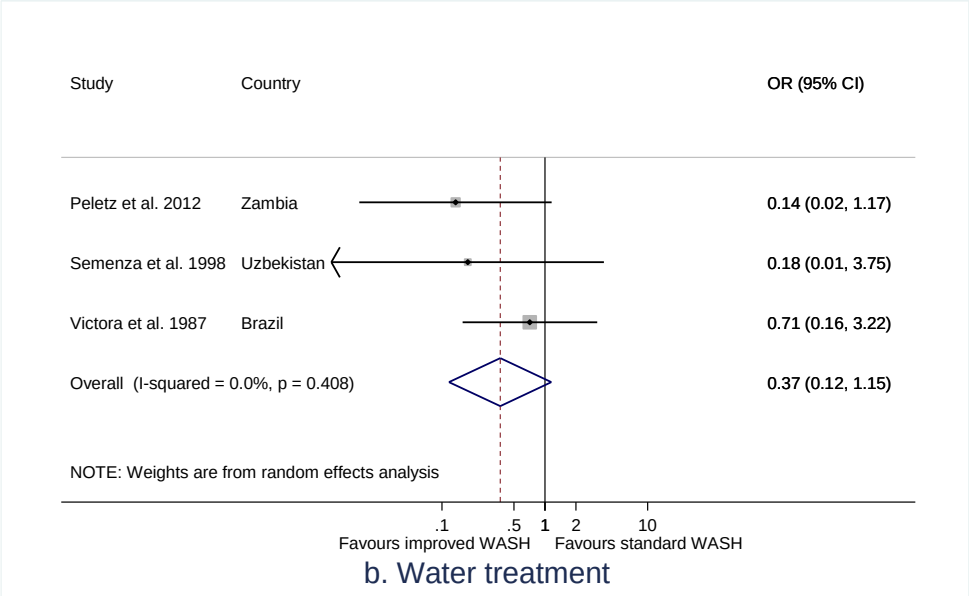
[156] The hygiene meta-analysis excluded the result from Bowen et al. (2012) because the water source, which was reporting as running for as little as two hours per week, would arguably not allow improved hygiene to be regularly practiced. When Bowen et al. (2012) was included, hygiene was associated with 56 percent reduction in diarrhoea mortality (OR=0.38; 95%CI=0.17, 0.86; 7 studies).

water treatment on diarrhoea mortality (OR=0.37, 95%CI=0.12, 1.15; 3 studies; p>0.1). Systematic review guidance does not give a minimum threshold on the number of studies that can be incorporated in a meta-analysis (Higgins et al., 2019), but test statistics such as I-squared are underpowered for small sample sizes (Higgins and Thompson, 2002). The limited number of water treatment or community latrine promotion studies that have examined mortality indicates the findings should therefore be interpreted cautiously.

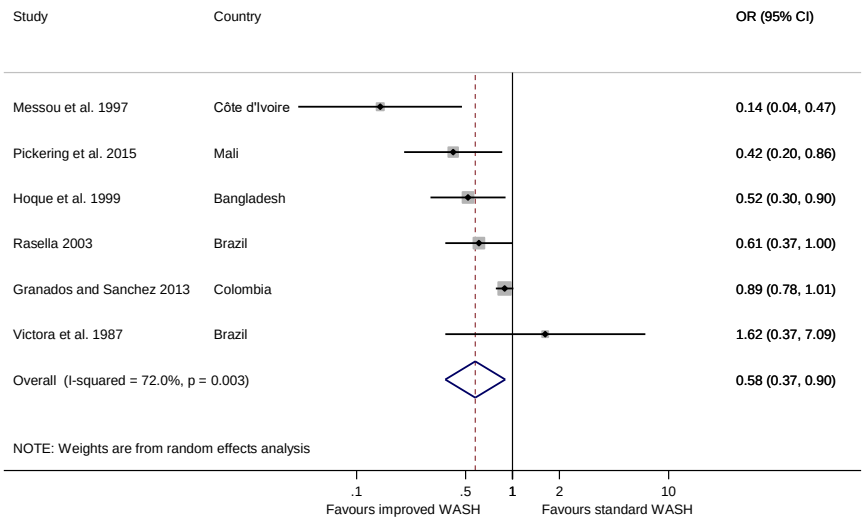
Figure 6.10 Diarrhoea mortality by WASH technology



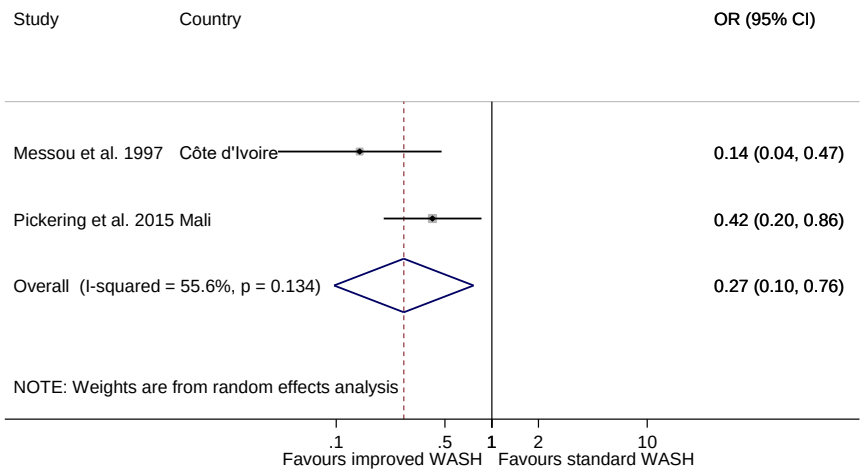
a. Water supply



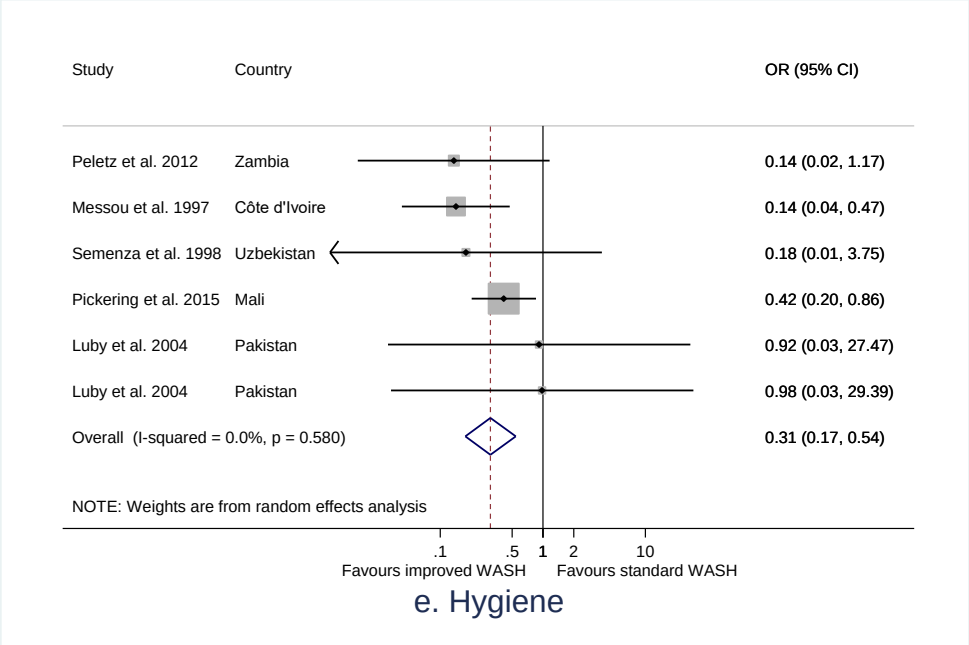
b. Water treatment



c. Household latrines



d. Latrines provided to entire community



e. Hygiene

A sensitivity analysis excluded one study thought to provide overestimates of water supply, sanitation and hygiene impacts on mortality, due to an ORS co-intervention (Messou et al., 1997) (Appendix D Figure D8). A second sensitivity analysis estimated pooled effects for the sample of intervention studies only (Appendix D Figure D9). The pooled effect from that analysis, indicating 61 percent reduction in diarrhoeal mortality due to of hygiene (OR=0.43; 95%CI=0.22, 0.84; 5 studies), was higher than the original estimate from Curtis and Cairncross (2003), which found 47 percent reduction associated with hand hygiene on diarrhoea morbidity. A recent systematic review of hygiene in schools also reported big reductions in diarrhoea disease of 53-73 percent (Mbakaya et al., 2017).

Secondly, multivariate meta-regression models were estimated for all-cause mortality (Table 6.6) and diarrhoea mortality (Table 6.7), to enable simultaneous examination of different competing sources of heterogeneity in findings across studies. For all-cause mortality, specification (1) was the least parsimonious. In specification (1), only length of follow-up and RCT design were associated with significant differences in all-cause mortality. However, owing to the large number of explanatory variables – including rural location, WASH technology provided, baseline water and sanitation, child's age, whether the child was immunocompromised, provision of ORS and health referrals and risk-of-bias status – the specification was underpowered to detect variation by other factors.

Specification (2) omitted the least significant background factors (rural, ORS, risk-of-bias rating),[157] and added an interaction of hygiene with baseline water supply to test for the water-washed route through which improved water increases chances of survival. The results indicated significant reductions in all-cause mortality when latrines were promoted to all households in community ($p<0.1$) but not when they were provided solely at the household level ($p>0.56$). There was no effect of hygiene provided in circumstances of unimproved water supply ($p>0.60$), although the effect was marginally insignificant when hygiene was given when water supply was also improved ($p>0.10$).

[157] Water treatment was included in specification (2) due to the policy interest in water treatment.

Table 6.6 Meta-regression analysis of all-cause mortality in childhood

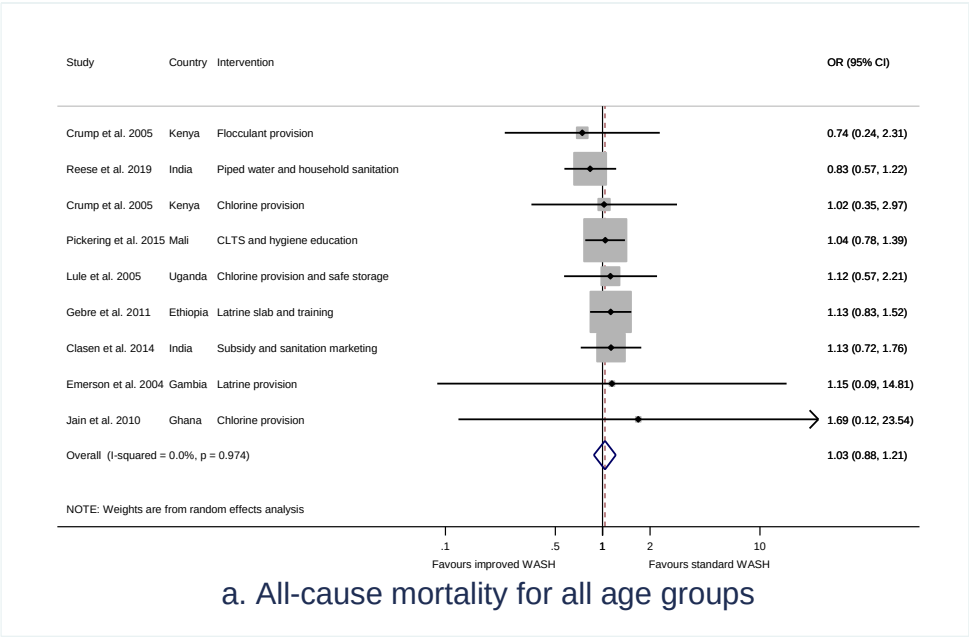
	(1)		(2)		(3)		(4)	
Regression variable	*OR*	*P>|z|*	*OR*	*P>|z|*	*OR*	*P>|z|*	*OR*	*P>|z|*
1=Rural	0.99	0.961						
1=Water supply	1.08	0.497	1.05	0.610				
1=Water treatment	0.99	0.980	0.99	0.966				
1=Household latrine	1.08	0.499	1.04	0.611				
1=Latrine to entire community	0.77	0.398	0.64	0.109	**0.72**	**0.081**	**0.72**	**0.087**
1=Hygiene	0.88	0.503	1.14	0.598				
1=Baseline water improved	0.89	0.102	0.91	0.140	0.95	0.285		
1=Hygiene with improved baseline water			**0.64**	**0.081**	**0.68**	**0.025**	**0.65**	**0.009**
1=Baseline sanitation unimproved	0.96	0.806	0.94	0.500				
1=neonates	1.11	0.313	1.13	0.188	1.14	0.165	1.12	0.198
1=infants	0.90	0.420	0.84	0.162	0.84	0.128	0.84	0.113
1=immunocompromised	0.44	0.396	0.33	0.156				
Follow-up (years)	**1.01**	**0.014**	**1.01**	**0.015**	**1.01**	**0.023**	**1.01**	**0.037**
1=ORS	0.96	0.942						
1=rainy season	0.29	0.122	**0.29**	**0.012**	**0.27**	**0.004**	**0.27**	**0.004**
1=RCT	**1.44**	**0.057**	**1.36**	**0.075**	**1.37**	**0.009**	**1.37**	**0.009**
1=High risk of bias	1.00	0.988						
Constant	0.83	0.340	0.93	0.663	0.92	0.289	0.93	0.339
Tau^2	0.012		0.009		0.007		0.006	
Residual I^2	65%		63%		62%		62%	
Adjusted R^2	-20%		15%		31%		36%	
Model F-test	1.24		1.77		2.77		3.07	
Num. obs	79		79		79		79	

Note: **bold** indicates coefficient is significant at p<0.1.

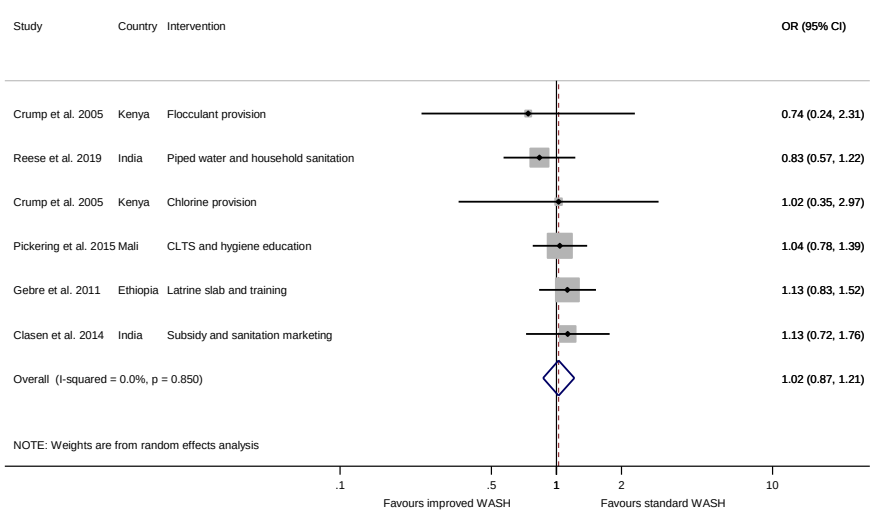
Specifications (3) and (4) omitted successive variables based on statistical significance, until the preferred specification (4) indicated significant effects on all-cause childhood mortality of hygiene when the baseline water supply was improved ($p<0.05$), and of sanitation when latrines were provided to the entire community ($p<0.1$). In addition, the effect on survival significantly increased in studies when the age group included infants ($p<0.1$), and when the study was done in the rainy season ($p<0.01$). Specification (4) also found smaller reductions in mortality when the study was an RCT ($p<0.01$) and when the follow-up period was longer ($p<0.05$).

The results broadly accorded with theory. For example, breast-feeding neonates would be exposed to limited faecal contamination from food or drinking water (Gautam et al., 2017), while weaning infants, who are constantly crawling on the floor and putting their fingers in their mouths, would be more susceptible to faecal-oral contamination. Furthermore, as noted by Butz el al. (1984), immunity systems mature with age, causing older children and adults to be less susceptible to infectious diseases. As a falsification exercise ('placebo test') of the findings for child mortality, meta-analyses were done of studies reporting mortality across the whole population and for those aged over 5 years. The findings do not suggest WASH provision leads to differences in all-cause mortality in all age groups (Figure 6.11a) or when restricted to adults and older children (Figure 6.11b).

Figure 6.11 Placebo tests



a. All-cause mortality for all age groups



b. All-cause mortality for adults and older children

The finding suggesting bigger effects of WASH provision in the tropical rainy season also accords with theory. Diarrhoea mortality in South Asia and sub-Saharan Africa has been shown as largely associated with *Escherichia Coli* infection in infants and cryptosporidium in children (Kotloff et al., 2013). Both are expected to be more prevalent in warmer conditions (Cairncross and Feachem, 2018). Since diarrhoea mortality is largely determined by verbal autopsy of carers, for which biases in reporting would not be expected to vary by season, this may also help to support the validity of the findings.

For child diarrhoea mortality (Table 6.7), column (1) presents results of eight meta-regressions including only a single explanatory variable and constant; they were estimated separately due to the very limited number of observations.[158] The remaining specifications tested for relevant relationships with limited numbers of explanatory variables. Specification (2) found no significant association between water treatment and mortality, while (3) and (4) tested the associations between diarrhoea mortality and transmission in, respectively, the public domain (community-wide sanitation) and household domain (domestic hygiene) (Cairncross et al., 1996). The findings suggested all variation in findings could be explained by four variables: whether community-wide latrines were provided or whether the intervention included a hygiene component; and study design and length of follow-up.

[158] No studies of diarrhoea mortality were limited to the rainy season, or among neonates and infants only.

Table 6.7 Meta-regression analysis of diarrhoea mortality

	(1)		(2)		(3)		(4)	
Regression variable	*OR*	*P>|z|*	*OR*	*P>|z|*	*OR*	*P>|z|*	*OR*	*P>|z|*
1=Water supply	1.08	0.824						
1=Water treatment	0.56	0.393	0.76	0.703				
1=Household latrine	0.93	0.833						
1=Latrine to entire community	**0.42**	**0.039**			**0.27**	**0.021**		
1=Hygiene	**0.50**	**0.067**					**0.22**	**0.034**
Follow-up (years)	**1.07**	**0.005**	**1.06**	**0.022**	**1.06**	**0.020**	**1.06**	**0.020**
1=RCT	0.76	0.545	0.91	0.820	**2.60**	**0.088**	**4.14**	**0.065**
1=High risk of bias	1.04	0.924						
Constant			**0.48**	**0.005**	**0.51**	**0.004**	**0.51**	**0.004**
Tau^2			0.000		0.000		0.000	
Residual I^2			24%		0%		0%	
Adjusted R^2			100%		100%		100%	
Model F-test			3.29		6.13		5.71	
Num. obs	17		17		17		17	

Notes: **bold** indicates coefficient significant at p<0.1. Column (1) presents bivariate meta-regression models where each coefficient is from a separate regression on the variable and a constant; (2)-(4) present multiple meta-regression analyses.

Finally, calculations were made of the prediction intervals (Chapter 4, Equation 4.13) for mortality associated with improved WASH. Unlike fixed effect meta-analysis, where the confidence interval of the pooled estimate incorporates the expected position of the treatment effect in a new context, the random effects estimate simply gives the mean across a range of distributions, each of which might contain the expected treatment effect in a new context. The prediction interval accounts for this additional uncertainty in the random effects estimator by providing a wider confidence interval that indicates the bounds on where the effect in a new context is likely to be. It aims to account for the between-study variance, which is larger where the pooled effect is estimated from fewer studies (Chapter 4, Section 4.2). The 95 percent prediction intervals (95%PIs) for most WASH technologies overlap the point of no effect (1), due to the limited numbers of estimates and therefore the large estimated between-study variance (Table 6.8). However, the findings also suggest that hygiene interventions, for which multiple consistent estimates were available, are likely to consistently reduce diarrhoea mortality in childhood when implemented in new contexts (OR=0.31; tau-squared=0.000; 95%PI=0.18, 0.55).

Table 6.8 Prediction intervals for random effects estimates

		OR	*SE*	*Tau-square*	*95% prediction interval*		*P>|z|*
All-cause mortality	Water supply	0.92	0.02	0.004	0.81	1.05	0.212
	Water treatment	0.88	0.14	0.000	0.67	1.16	0.375
	Sanitation	0.91	0.03	0.007	0.76	1.09	0.313
	Community sanitation	0.83	0.15	0.000	0.62	1.10	0.268
	Hygiene	**0.86**	**0.09**	**0.000**	**0.72**	**1.02**	**0.085**
Diarrhoea mortality	Water supply	0.66	0.15	0.070	0.36	1.20	0.204
	Water treatment	0.37	0.58	0.000	0.12	1.15	0.228
	Sanitation	0.58	0.22	0.177	0.23	1.47	0.293
	Community sanitation	0.27	0.53	0.323	0.06	1.24	0.235
	Hygiene	**0.31**	**0.29**	**0.000**	**0.18**	**0.55**	**0.010**

Note: **bold** indicates coefficient is significant at P<0.1; *SE* is the natural logarithm of the standard error of *OR*.

6.7 Discussion and implications

These findings are remarkably consistent with theoretical predictions. First, one would expect a stronger relationship between improved WASH access

and diarrhoea mortality, than all-cause mortality. This is borne out by the estimated effect of improved WASH on diarrhoea mortality of around 30 percent, as compared to 10 percent reduction in all-cause mortality. Inadequate WASH may cause death in young children through other routes such as respiratory infection, under-nutrition and even safety of the WASH technology itself,[159] but diarrhoea is by far the biggest cause (Prüss-Üstün et al., 2019). Hence one would expect a bigger reduction in diarrhoea mortality over a smaller denominator.

The significantly bigger effects of community-wide sanitation interventions, and hygiene over other WASH technologies, are important findings. This evidence suggests that the crucial factors in combating death in early childhood in L&MICs are hygiene promotion, which is most likely to operate in the household domain, and community-wide sanitation, which reduces open defaecation in the public domain (household sanitation provision was not correlated with mortality). From the meta-analysis of diarrhoea mortality, three of the five biggest effects were from studies of multiple WASH technologies with a hygiene component in Côte d'Ivoire (Messou et al., 1997), Uzbekistan (Semenza et al., 1998) and, alongside CLTS, in Mali (Pickering et al., 2015). The fourth and sixth biggest were of piped water provision in Brazil (Victora et al., 1988) and Uzbekistan (Semenza et al., 1998).

The analysis also suggested a mechanism through which water affects mortality, by enabling hygienic practices around handwashing, food preparation and cleanliness in the household (fomites). Effects in individual studies of hygiene also appeared related to water supply access. For example, in Messou et al. (1997), hygiene education was provided alongside village water pumps which gave 76 cubic metres per day for a community of 400 people, equivalent to 190 litres per capita per day. The study with smallest effect on diarrhoea mortality was conducted among communities where some households had access to running water for only two hours each week (Bowen et al., 2012).[160] These findings are consistent with water-washed

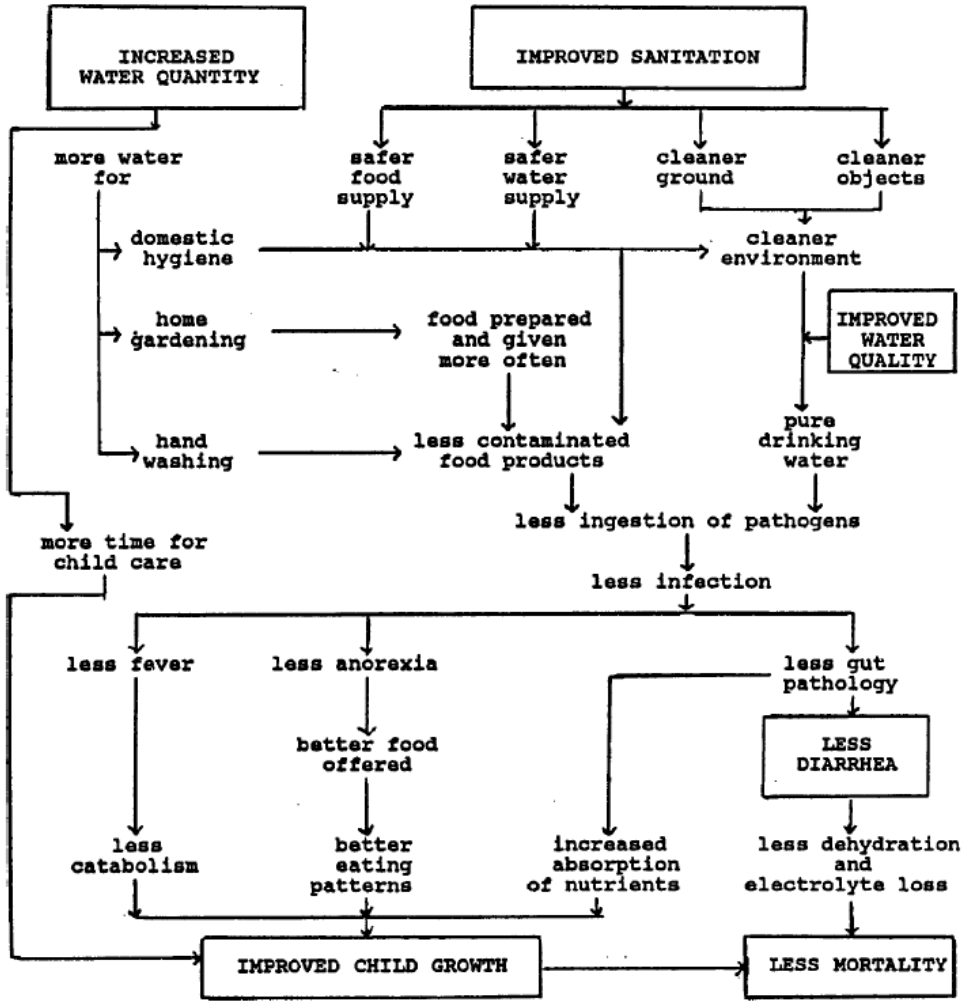
[159] In addition to drowning as a source of death, there are also reports of accidents and one death in Bangladesh due to a child or elderly person falling into a latrine pit, because the latrine slab was not made of concrete that was not reinforced (Hanchett et al., 2011).

[160] Shrestha et al. (2013) also found greater water consumption of any quality to be more effective in addressing diarrhoea morbidity in Nepal than limited water of better quality, which they related to personal hygiene.

faeco-oral infection being the main cause of childhood mortality in endemic circumstances (Cairncross and Feachem, 2018).

The corollary of this is the finding that water treatment does not appear to be an effective means of combating mortality in endemic circumstances. This suggests that the principal transmission route of faeco-oral infection is not usually water-borne. The F-diagram includes six intermediate transmission vectors (fluids, fields, flies, fingers, food and fomites), of which only the fluids route is addressed through water quality. Esrey (1987) presented a logic model showing the theoretical relationship between water supply, water treatment, sanitation and hygiene, on the one hand, and diarrhoeal disease, child nutritional status and survival, on the other (Figure 6.12).

Figure 6.12 Relationship of improved water, sanitation and hygiene to diarrhoea, child growth and mortality among young children



Source: Esrey (1987).

The figure indicates that the routes from water supply and sanitation to survival operate through various intermediate quality of life outcomes relating to better hygiene practices (including hand and food hygiene, and 'fomites') and childcare, diarrhoeal disease and nutrition. Many of the papers included in the meta-analysis primarily analyse the effects of WASH on these intermediate outcomes. Mediator analysis could therefore be done to explore the relationships between these outcomes and survival, to shed further light on the routes to improving survival.

Evidence suggests that pathogens in food may be a much more important source of faeco-oral disease than those in drinking water. Motarjemi et al. (1993) reviewed theory and evidence on possible contamination of weaning foods from fields, flies, fingers and fluids. For example, Barrell and Rowland (1980) found that gruels used as weaning food in the Gambia contained 100 times more *E. Coli* per 100 ml after two hours than the water used to prepare them, and 10,000 times more after eight hours. In contrast, diarrhoeal risk may only increase in drinking water where *E. Coli* contamination exceeds 1,000 faecal coliforms per 100 ml (Moe et al., 1991). More recently, bacterial contamination of weaning food was high in almost 90 percent of stored weaning food in Mali (Touré et al., 2013), and in Zambia, half of samples were contaminated with *E. Coli* or Salmonella (Kinkese et al., 2018).

The greater effects of hygiene on mortality when water supply is improved, and of sanitation when provided community-wide, but not individually, is also consistent with a hypothesis of threshold effects (Shuval et al., 1981), as also shown recently in meta-analysis by Wolf et al. (2019). Under this hypothesis, environmental pathogen exposure in environments with unimproved water and sanitation is sufficiently great, that household provision of WASH technology may an ineffective means of combating infectious disease mortality. Where community-wide sanitation and water conditions are improved, hygiene is more effective in improving child survival, since there is less faecal matter in the public domain, and improved water supply enables adequate hand and food hygiene in the private domain (Cairncross et al., 1996).

The evidence presented in this Chapter suggests important findings for the sequencing of WASH technology improvements if the primary aim is to stop children from dying before the age of 5. Where people have access to

improved water supplies, hygiene promotion may be able to combat infectious disease mortality in the domestic domain, where it is likely to be greatest in early childhood. Sanitation makes a difference for child survival when latrine provision is community-wide. It is known that water supplies are a pro-poor and gender-inclusive intervention due to the time-savings they enable (Cairncross and Cliff, 1987; Churchill et al., 1987). These results suggest that improved water supplies, in combination with hygiene and community-wide sanitation promotion, should also be prioritised for the potentially vast impacts on child survival.

A final analysis was done to update the estimates of diarrhoeal deaths in the global burden of disease due to inadequate WASH by the WHO (Prüss-Üstün et al., 2019). The calculations here used the same approach and dataset, calculating the disease risk using the population attributable fraction method (Vander Hoorn et al., 2004; Lim et al., 2012). Two calculations were made, which were deemed to provide lower- and upper-bound estimates of mortality (Table 6.9). Lower-bound estimates were from the sensitivity analysis of diarrhoea mortality presented in Appendix D Figure D8. Upper-bound estimates were from Figure 6.10.

The lower-bound findings indicated that, presently, GBD diarrhoeal deaths due to WASH may be underestimated by nearly half a million people, most of whom live in WHO Africa region. Furthermore, half of all diarrhoeal deaths, almost 700,000 people, may be caused by inadequate hygiene. The upper-bound estimates suggested GBD underestimates diarrhoea deaths due to WASH by nearly two-thirds, including an additional half a million deaths due to inadequate sanitation coverage, and half a million due to hygiene and water supply. Nearly all of the extra diarrhoeal deaths are in sub-Saharan Africa. These results call for a reprioritisation of resources to Africa, and for interventions promoting safe hygiene practices and sanitation coverage.

The finding concurs with a previous review of diarrhoea morbidity by Curtis and Cairncross (2003), who stated: "…current evidence shows a clear and consistent pattern. If handwashing with soap could save over a million lives, if rates of handwashing are currently very low, and if carefully designed handwashing promotion programmes can be effective and cost-effective, then handwashing promotion may become an intervention of choice" (p.280).

Table 6.9 Diarrhoeal disease deaths due to inadequate WASH

Bound	*WHO region*	*Water*	*Sanitation*	*Hygiene*	*Total*	*Total (WHO)*
Lower	Europe	479	412	533	1,425	1,500
	Western Pacific	5,207	6,783	8,540	20,530	11,600
	Americas	1,549	3,173	7,211	11,932	9,800
	Eastern Med.	16,318	19,915	37,485	73,718	76,300
	South East Asia	63,443	84,750	114,006	262,200	295,100
	Africa region	176,068	247,995	497,134	921,197	431,700
	Total	263,064	363,028	664,910	1,291,001	826,000
Upper	Europe	785	1,830	786	3,401	1,500
	Western Pacific	8,219	23,111	11,947	43,277	11,600
	Americas	2,485	11,372	9,843	23,700	9,800
	Eastern Med.	25,076	56,768	47,843	129,687	76,300
	South East Asia	98,317	240,016	151,939	490,273	295,100
	Africa region	266,626	619,187	593,732	1,479,545	431,700
	Total	401,508	952,285	816,090	2,169,883	826,000

Note: WHO regions used. Total (WHO) are the estimates from Prüss-Üstün et al. (2019).

Source: author.

Chapter 7 Conclusion: getting WASH impact evaluation right from the bottom up

7.1 Introduction

The Thesis has examined bias in impact evaluations. The main focus has been on water, sanitation and hygiene, for which access and use is fundamental for survival chances in childhood, basic needs like nutrition, excretion and safety, and higher order needs like dignity, productivity, and happiness. This chapter overviews the main findings, with respect to the Thesis questions posed in Chapter 2, and discusses their implications. Section 7.2 presents the contributions of the Thesis to answering the research questions, and the limitations of the work done. Section 7.3 presents conclusions for policy and future research.

7.2 Findings and limitations of this Thesis

The importance of WASH is recognised in the increased funding and attention devoted for global policy and programmes and to enabling rigorous research about what works and why. Such questions are increasing answered using RCTs. Well-conducted RCTs are usually favoured to answer causal questions. However, there are important concerns about bias due to problems in design and implementation, making RCTs potentially no more reliable than non-randomised studies. Moreover, prospective studies including RCTs cannot usually assess adequately important questions for policy like impacts of WASH on child survival, due to statistical power and ethical reasons. There has therefore been a simultaneous rise in the production of systematic reviews, which aim to address these problems by drawing on evidence from multiple studies.

To answer Thesis Question 1 on the types of interventions, outcomes and study designs covered in WASH impact evaluation and systematic reviews, Chapter 3 reported a big increase in production of both types of studies since the International Year of Sanitation 2008. This corresponded to a 'behavioural revolution' in policy research, where the focus has increasingly

shifted from evaluating WASH technology provision to WASH promotional approaches to incentivise uptake and adherence. Resources have also become available for multiple and large-scale intervention studies using RCTs and prospective non-randomised (quasi-experimental) approaches, of which an estimated 350 in L&MICs have now been reported.

Chapter 3 also addressed Thesis Question 1 on rigour, relevance and representation in WASH intervention research, finding that quality standards have improved, but also concerns about the ways in which research resources are distributed and primary studies and evidence syntheses routinely done. For example, findings indicated limited correlation between important outcomes for stakeholders, whether priorities are set from the top down or the bottom up, and research priorities, measured by numbers of impact evaluations or participants in those studies (although the correlation within outcomes by geographical distribution of studies and disease burden was high). Most impact evaluations and reviews have been led by researchers based at academic institutions in Western countries, and it is not clear to what extent researchers from L&MICs are involved substantively in study design and analysis. There is a risk that the questions answered will not reflect local priorities or, in particular, not be taken up by policymakers in the contexts where the studies are based. This distribution also distorts views about WASH impact evaluation research, when the first impact evaluations of WASH in L&MICs were done by researchers based in Bangladesh, Brazil, Guatemala and Mozambique. It is slightly different for systematic reviews, as these efforts have tended to be led in high-income countries, although efforts are being made to change that.

There is also a wealth of information from research that was not eligible for inclusion in the census of WASH intervention studies. For example, evidence was omitted from low-income contexts in high-income countries, which may be comparable to, and therefore provide relevant evidence for, L&MIC contexts (Rosling et al., 2018). In addition, studies were excluded that only presented evidence on intervention processes or participant views (studies that did not contain, or require, strong counterfactuals), economic evaluations or impact evaluations examining the relationship between exposures and outcomes, without clear reference to an intervention. This limitation was partially rectified by incorporating non-intervention exposure studies into the analysis of child survival in Chapter 6.

There is much greater scope in WASH research for using credible non-randomised approaches to answer pressing questions that RCTs cannot, such as to provide evidence on survival or long-term effects of interventions or exposures. These questions can be answered using natural experiments – causal studies conducted retrospectively using existing data (e.g., household surveys or administrative records) with selection on unobservables. While observational studies are more likely to subject to confounding bias than RCTs, they may be less at risk of other bias including departures from intended interventions due to motivation bias (e.g., Hawthorne effects). Natural experiments (e.g., regression discontinuity designs, RDDs) on the other hand, can estimate an unbiased causal effect in expectation (i.e., they can account for unobservable confounding), without risk of motivation bias. However, these studies may be subject to sampling bias in estimating the population treatment effect and, when inappropriately designed or executed, may be subject to other biases. The conduct of these studies must necessarily incorporate confirmation and falsification exercises to support statistical inferences.

Chapter 4 developed a critical appraisal tool to assess bias transparently and consistently across bias domains for RCTs and non-randomised studies, including natural experiments, addressing Thesis Question 2. 'Signalling questions' were incorporated to evaluate specific biases for non-randomised studies with selection on unobservables such as RDDs, difference studies and instrumental variables. Signalling questions for performance and motivation bias were also developed, as these were insufficiently articulated in existing risk-of-bias tools, including those designed to evaluate RCTs.

Some tests of the risk-of-bias tool were presented based on systematic reviews conducted during the Thesis period. Two researchers working independently to assess risk of bias were able to reach agreement about scores in all areas except performance and motivation bias, where expected agreements were below those expected by chance in one pilot exercise (i.e., Cohen's kappa ≤ 0). Factors relating to the intervention affect risk of performance and motivation bias, such as whether it is delivered in the form of information that can 'spillover' to controls, whether controls can crossover to obtain treatment – and therefore whether geographical separation is necessary and sufficient (and if so, how far away from one another they

should be) – or whether repeated observation might reasonably affect adherence among treated units (or adherence due to information provided to controls through 'survey effects'). Clearly articulated signalling questions for these sources of bias need to be incorporated into critical appraisal. The pilot review where agreement could not be reached about performance and motivation bias comprised a variety of interventions and outcomes, making precise questions difficult to articulate and therefore increasing the role of reviewer judgement in reaching bias decisions. This limitation was addressed in the systematic review reported in Chapter 6, which clearly articulated signalling questions to evaluate WASH-related mortality.

Thesis Question 3, which asked whether the biases predicted in theory are borne out empirically, was answered in Chapter 5 drawing on internal and external study replications. The chapter examined the circumstances in which non-randomised studies produced the same estimated effects as RCTs. In the first part, statistical meta-analysis was used to synthesise pooled effects from 17 systematic reviews and meta-analyses across various topics in international development (e.g., agriculture, climate change, economic development, education, governance) which had themselves used the critical appraisal approach discussed in Chapter 4. Focusing on the relationship between predicted bias, and the distribution of pooled effect sizes obtained from random effect meta-analysis, using external replications – that is, studies assessing the same intervention and outcomes in different contexts and target populations – the results indicated that relatively well-conducted NRS, including those with 'low risk of bias' or 'some concerns', estimated the same pooled effects on average as RCTs across 39 comparisons. In other words, the average difference *D* between standardised pooled effects was found to be zero (D=0.00; 95%CI=-0.06, 0.06) when comparing 'low risk' NRS with RCTs and indistinguishable from zero (D=0.01; 95%CI=-0.03, 0.05) when comparing NRS with 'some concerns' and RCTs. Where NRS are eligible for inclusion in systematic reviews, it is usually justified for external validity; this analysis suggests another reason, namely that well-designed and implemented NRS also provide internally valid effect estimates.

However, 'high risk' NRS on average estimated significantly bigger pooled effects (D=0.17; 95%CI=0.07, 0.28), demonstrating why risk-of-bias assessment is a key component of meta-analyses of such studies. Whereas NRS with greater risk of bias on average produced effects of greater

magnitude, the analysis suggested that RCTs with greater risk of bias produced effects of significantly lower magnitude (D=-0.08; 95%CI=-0.14, -0.03). Well-implemented RCTs may therefore have other attributes, such as being located in favourable contexts or having more careful intervention fidelity, which can lead to larger effects.

All of the findings were robust to sensitivity analysis where 'pooled effects' comprising only a single study (or two studies) were excluded from estimation. A limitation of the analysis is that the included studies were not found through systematic searches, but rather opportunistically, as the reviews that had used the approach developed by the author in his capacity as Editor of the Campbell Collaboration Coordinating Group which supported the reviews. So, while there is high confidence that the findings are representative of the population of systematic reviews in international development that used the risk-of-bias approach, further synthesis research is needed to assess whether the findings are representative more broadly.

The second part of Chapter 5 synthesised evidence from a systematic review of internal replication studies in international development – that is, studies that, for the same context and target population, compare the results of a NRS estimate with a benchmark estimate from a well-conducted randomised study. Using fixed-effect meta-analysis to synthesise the evidence, internal replications using selection on unobservables produced estimates that were almost identical to RCTs, including RDD (mean squared error=0.00), credible instrumental variables (MSE=0.00) and double differences (MSE=0.02). Studies with selection on observables, such as statistical matching, produced effects that more closely approximated those from the RCT when incorporating baseline outcomes (MSE=0.00), and geographically local matches (MSE=0.03). A key implication of the analysis is that rigorous studies using natural experimental approaches, which are an underutilised approach in WASH impact evaluation, can provide unbiased estimates where randomisation is not feasible or ethical.

Many systematic reviews and meta-analyses have been conducted to synthesise findings on the effects of WASH technology provision on infectious diseases, usually diarrhoea morbidity. But the underlying assumption of these analyses is that diarrhoea morbidity is a good proxy for diarrhoea mortality, which is the biggest component of the global disease

burden relating to inadequate WASH. There is no existing systematic review of child mortality data outcomes due to WASH, despite the large number of observational NRS estimating the relationship. Furthermore, child mortality can be estimated from RCTs by synthesising data from participant flow diagrams in trials. Drawing on studies in the WASH intervention evidence census that reported mortality, together with studies examining exposures that were excluded at full-text stage in Chapter 3, and critically appraising these studies using the modified tool from Chapter 4, Chapter 6 addressed Thesis Question 4 by estimating the effects of WASH provision on all-cause and infectious disease-related mortality.

No studies were found to have low risk of bias in estimating effects of WASH on mortality, and RCTs with high risk of bias were found to have smaller effects than other RCTs, echoing the findings in Chapter 5. Publication bias analysis suggested that there was no evidence of small-study effects among prospective intervention studies including RCTs reporting mortality, precisely because mortality was not usually a primary study outcome (or, indeed, defined as an outcome at all where the mortality estimate was taken from the participant flow diagram). It is only rarely that formal publication bias does not find any evidence for small study effects (Rothstein et al., 2005). In contrast, evidence on small study for non-randomised studies, which largely reported mortality as the primary outcome, suggested the presence of publication bias.

Whereas only a single intervention study was able to report a statistically significant effect, the meta-analysis results indicated that WASH provision and promotion at the household level led to approximately 15 percent reduction (OR=0.87; 95%CI=0.77, 0.98) in all-cause mortality in childhood, and over 50 percent fewer child diarrhoea deaths (OR=0.44; 95%CI=0.24, 0.80), relative to control communities. Further analysis indicated that the statistical heterogeneity in reductions in childhood diarrhoea mortality across studies was explained by two sets of variables: hygiene in the public and household domains, as measured by interventions promoting community-wide sanitation and domestic hygiene; age of child where impacts were bigger among post-neonatal infants, who were more likely to be weaning and have weaker immunity than older children; and study design, where RCTs systematically found smaller, but still beneficial, effects on mortality than non-randomised studies (as would be expected when

comparing intention-to-treat analysis used in RCTs with treatment-on-the-treated analysis in NRS).

The main limitations of the analysis in Chapter 6 are that it draws largely on an evidence census conducted in 2018 – hence the searches are outdated – and there were limited attempts to contact authors to obtain unpublished information. Efforts were made in 2020 to locate completed reports of evaluations that had been registered by the time that searching was conducted in 2018. In addition, although some authors were contacted, comprehensive efforts to obtain unpublished information and datasets would enable fuller analysis of the available evidence on all-cause mortality.

7.3 Implications for policy and further research

The findings in this Thesis suggest an important role for hygiene in combating death in childhood, particularly in sub-Saharan Africa. In Chapter 6, reductions in mortality were found to be significantly higher when WASH interventions included domestic and public hygiene components, and hygiene interventions were also more effective in combating mortality when water supply access was more reliable. These findings also suggest water supply is an important enabler of domestic hygiene, by acting on the quantity of water available for use. It is possible that the mechanisms through which water supply's effectiveness on diarrhoea operate are dependent on the context, including whether the situation is endemic or epidemic, as well as factors like distance to the source, reliability of supply and cultural factors determining when weaning begins, for example. However, there were no estimated effects on all-cause or diarrhoea mortality in childhood of water treatment, which acts on water quality. This result supports the notion that faeco-oral infection in endemic conditions is transmitted primarily through the water-washed route, owing to inadequate hygiene and water supply, including for weaning infants and young children. The implications of the findings for the WHO's Global Burden of Disease are substantial. Diarrhoea mortality may be under-estimated by at least half a million people every year, and possibly as much as 1.3 million, mainly people living in sub-Saharan Africa.

In the area of interventions, where faddism can easily take root, the Thesis suggests that rigorous evidence can support decision making, by providing

contextually relevant and generalisable evidence, and the apparatus to distinguish between the two. Rigorous evidence about interventions operating to stimulate both demand for WASH technologies and supply, including for improved performance of WASH institutions, are thin on the ground. For example, nearly all of the studies of decentralisation are of a single approach, community-driven development (CDD), but there is as yet no systematic, critically-appraised evidence on WASH benefits due to CDD or other forms of decentralised service delivery.

The Thesis has also shown that rigorous causal evidence can be obtained in contexts where randomisation is not feasible (Chapter 5). There are more opportunities to conduct rigorous observational studies with 'as-if' randomisation than are presently taken in WASH impact research, particularly natural experiments using regression discontinuity design, interrupted time-series of administrative data, and other approaches. This is an area where rigorous, relevant and cost-effective studies in WASH could proliferate, as they have done elsewhere. There could also be more recycling of existing data, whether through new primary studies based on administrative data, or new syntheses such as that presented in Chapter 6, than is presently done.

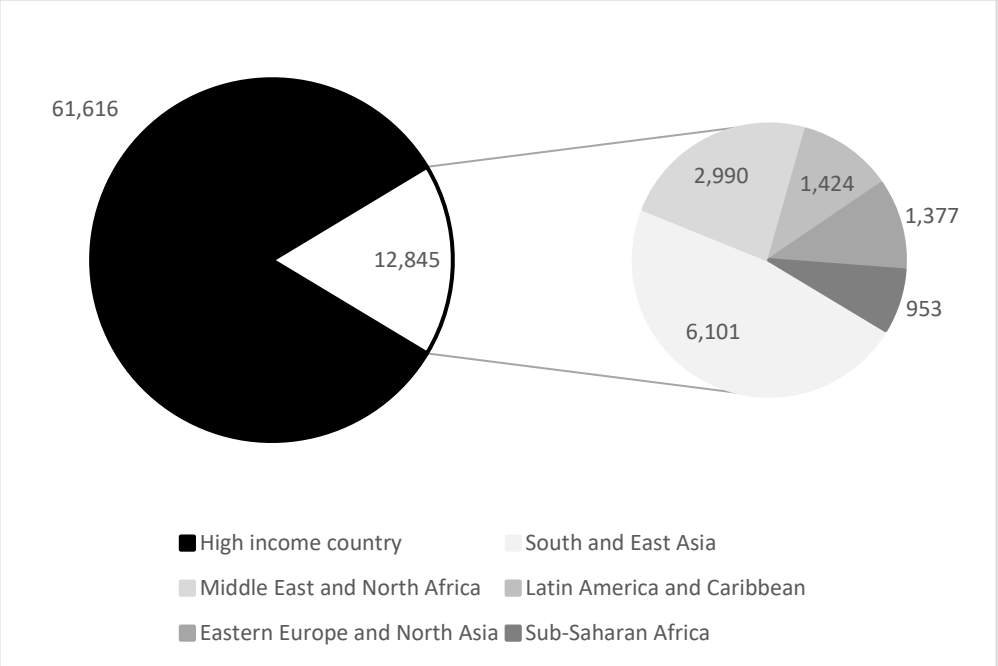
Efforts should be refocused to evaluate, and synthesise evaluations of, outcomes of importance for alleviating the global disease burden, especially reduced respiratory infection from handwashing (only 34 studies over a population of 125,000 participants, although this has very likely changed due to the COVID19 pandemic), musculoskeletal disorders from water-carrying (a single study with 2,500 participants) and pedestrian road injuries (no studies). They should also include outcomes of importance to people with poor WASH access, especially fears about safety and psychosocial stress (only four studies with 4,500 participants). There is a particular need for studies, and synthesis of studies, examining the impacts of water supply and sanitation improvements on women's time use using rigorous observational approaches, and the incorporation of the time use as standard in household surveys.

The risk-of-bias approach developed in Chapter 4 and piloted in Chapters 5 and 6 shows that critical appraisal can be operationalised to assess randomised and non-randomised studies transparently across the same bias

domains. This is consistent with initiatives that draw on critically appraised evidence from randomised and non-randomised approaches to assess the body of evidence, in particular GRADE. However, these approaches depend on the quality of reporting in primary studies. The greater opportunities online publication affords for reporting of supplementary information facilitates transparent reporting of research conduct and findings.

Clear progress could be made towards improved registration and reporting of non-randomised studies. For example, there are around 75,000 observational studies registered with clinicaltrials.gov, of which 20 percent are conducted among L&MIC populations (Figure 7.1). 3ie's registry of studies in L&MICs includes 80 studies out of 202 that use non-randomised approaches, and 27 randomised and non-randomised studies in the WASH sector alone.[161] Only one WASH impact study was pre-registered (Reese et al., 2017), indicating it is possible to pre-register analyses, and publish findings without precisely estimated effects, even with retrospectively designed evaluations.

Figure 7.1 Number of NRS clinical trials registered by region



Source: author using data from clinicaltrials.gov.

[161] In addition, of the 1,628 studies registered with EGAP, 15 were identifiable as NRS (double differences, statistical matching, natural experiments, regression discontinuity design, interrupted time series) of which eight were in L&MIC populations (searches undertaken on 3 September 2020).

Duflo et al. (2020) recently made a case 'in praise of moderation' of the use of pre-analysis plans (PAPs) – a paean to a more minimal approach to the pre-registration of development economics field trials.[162] This supported the current status quo of pre-registration "to the extent possible" prior to intervention start in the American Economic Association (AEA) RCT registry (which two of the authors had established), but arguing that anything more burdensome may cause researchers to "be discouraged from looking at outcomes which are important but imprecise and self-censor the set of ideas they pursue" (p.5), thus stifling scientific progress.

There are several reasons why these concerns are overstated. As argued by the Executive Director of Evidence in Governance and Politics (EGAP) research network, Duflo et al. (2020) seem to underestimate the value of the PAP, not as a document but as a process to obtain crucial feedback from stakeholders, so help to avoid wasting opportunities (or 'messing up') expensive evaluations.[163] This is particularly important as the opportunity costs of evaluation resources are hardly negligible, and include providing capacity building in L&MICs to undertake good research. Secondly, Duflo et al. (2020) over-state the binds that PAPs place on researchers' choices in analysis. In the systematic reviews community, for example, pre-analysis plans (called protocols) are registered and peer reviewed as standard, including in international development. Methods will ideally be pre-specified as far as possible during the study design phase (e.g., sources of data, methods of synthesis, moderator and sub-group analyses), by drawing on programme theory, and feedback from a review advisory group. This is to minimise bias, or perceptions thereof, in the research process. It is therefore a requirement of Cochrane and Campbell Collaboration reviews that moderator variables and discussion of potential moderator analyses are presented in the study protocol. However, it is reasonable to expect some analyses to be identified post hoc, and it is therefore common for studies to deviate from protocol. This is accepted practice, which reporting standards

[162] Although several other registries open to non-randomised studies are listed, including 3ie's Registry for International Development Impact Evaluations https://ridie.3ieimpact.org/ and EGAP's registry http://egap.org/content/registration, Duflo et al. (2020) argue that pre-registration "for non-experimental research, which tends to be retrospective, are rarely advocated for or used (yet) in practice, presumably because they are neither desirable nor, in most cases, practical" (p.4).

[163] Cyrus Samii, "Using pre-analysis plans to learn better and to learn together", 21 April 2020: https://cyrussamii.com/?p=3154 (accessed 21 July 2020).

allow for, provided deviations are transparently indicated.[164] For example, moderators and sub-groups may be identified during the data collection phase, some analysis variables may be open-coded qualitatively, and subsequently grouped into quantitative codes, or it may be a component of certain types of study, for example, where mixed methods are used to integrate the findings from syntheses of quantitative and qualitative data, which may necessitate an iterative approach to data collection and analysis. Similar approaches could be adopted for primary studies, whether randomised or prospective non-randomised studies, or retrospective studies (natural experiments and observational studies).

Thirdly, while study designs have clearly improved over time, Chapter 3 showed that transparency in reporting is extremely weak. Although standards for impact evaluation design have improved over time, fewer than half of WASH trials in environmental health presented participant flow diagrams, and less than 5 percent have done so in the social sciences. By far the most obvious improvement that can be made for trials (prospective randomised or non-randomised studies), therefore, is for authors to report (and journals to require publication of) full participant flow diagrams according to accepted standards following CONSORT and its adaptation for the social sciences (e.g., Bose, 2010). These diagrams should clearly indicate the sequencing of participant recruitment in relation to cluster-randomisation, losses to follow-up, and reasons given for losses, including permanent migration and death. In addition, clarity is needed on the methods used to randomise participants and conceal allocation until recruitment. Although it is not usually possible to blind participants to interventions, it is possible to blind data analysts to intervention status, or outcome assessors in cluster trials. It is also possible to obtain informed consent without clearly linking data collection to the intervention, effectively blinding participants to the trial, reducing risk of courtesy bias, and reassure respondents that answers are not going to be used to determine further assistance or project.

In order to have the biggest effect on improving the lives of people who participate in these studies and those targeted by the interventions they are

[164] Methodological Expectations of Campbell Collaboration Intervention Reviews (MECCIR): https://campbellcollaboration.org/about-meccir.html (accessed 21 July 2020).

evaluating, the culture of evaluation publishing should shift towards the transparency promoted by approaches like the Nakuru Accord (Box 7.1), developed at the Water Engineering Development Centre Conference (WEDCC) in 2018 to provide a set of ethical research principles. According to the website, it has been signed by over 250 individual researchers and 12 organisations.[165] Signatories commit to being transparent about 'failures', which in the area of impact evaluation research may, *inter alia*, be closely related to reporting findings regardless of their statistical significance and greater transparency in decision-making about specification searches.

Box 7.1 The Nakuru Accord: failing better in the WASH sector

Transparency and accountability are necessary for achieving sustainable, positive impacts from water, sanitation and hygiene. As a WASH professional, I believe that we can achieve this through a culture of sharing and adaptation when things go wrong. To support this, I will:

- Promote a culture of sharing and learning that allows people to talk openly when things go wrong.
- Be fiercely transparent and hold myself accountable for my thinking, communication and action.
- Build flexibility into funding requests to allow for adaptation.
- Design long-term monitoring and evaluation that allows sustainability to be assessed.
- Design in sustainability by considering the whole life cycle.
- Actively seek feedback from all stakeholders, particularly end-users.
- Recognise that things go wrong, and willingly share these experiences, including information about contributing factors and possible solutions, in a productive way.
- Critically examine available evidence, recognising that not all evidence is created equal.
- Write and speak in plain language, especially when discussing what has gone wrong.

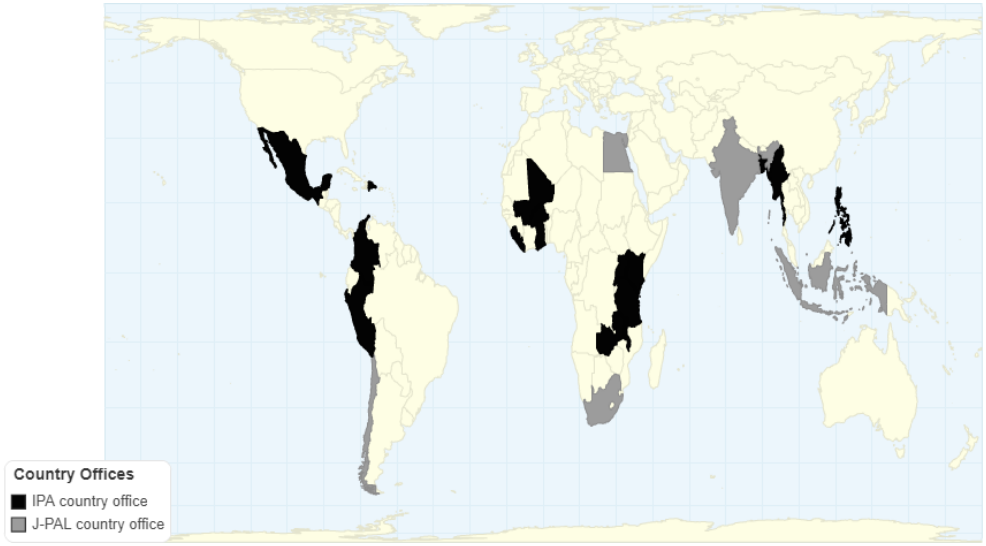
Source: https://wash.leeds.ac.uk/failing-better-in-the-wash-sector/.

Progress could also be made to address the imbalance in global research resources, towards L&MIC institutions and HIC institutions with close partnerships with L&MICs. Since 2000, with the rise of the 'big three'

[165] https://wash.leeds.ac.uk/failing-better-in-the-wash-sector/ (accessed 15 September 2020).

producers of WASH development impact evaluations – the Abdul Latif Jameel Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA) and the World Bank – the research capacity to conduct RCTs in L&MICs has increased considerably. 3ie estimated that there were over 4,000 development impact evaluations across sectors (Sabet and Brown, 2018). The evidence census found over 350 studies of WASH interventions in L&MICs, of which over half were RCTs, mostly published since 2008. J-PAL and IPA have also established country offices specialising in implementing RCTs and collecting the survey data on which they are analysed (Figure 7.2). With the establishment of capacity building initiatives like the Centers for Learning on Evaluation and Results (CLEAR),[166] where J-PAL South Asia is based, and organisations like 3ie, whose first office was established in New Delhi, and which encouraged authorship by favourably weighting scores for teams involving L&MIC staff in meaningful research roles (study design, data analysis, writing up) on grant applications, one would therefore expect there to have been substantial opportunities for the building of leadership and research capacity in impact evaluation in L&MICs.

Figure 7.2 Location of J-PAL and IPA country offices[167]



Source: chartsbin.com.

[166] https://www.theclearinitiative.org/ (accessed 15 September 2020).
[167] IPA's locations are: Nairobi, Kenya; Lilongwe, Malawi; Kigali, Rwanda; Dar es Salaam, Tanzania; Kampala, Uganda; Lusaka, Zambia; Ouagadougou, Burkina Faso; Abidjan, Côte d'Ivoire; Accra, Ghana; Tamale, Ghana; Monrovia, Liberia; Bamako, Mali; Freetown, Sierra Leone; Dhaka, Bangladesh; Yangon, Myanmar; Pasig City, Philippines; Bogotá, Colombia; Santo Domingo, Dominican Republic; Mexico City, Mexico; Lima, Peru; Washington, DC, USA. J-PAL's regional offices are as follows: Cairo, Egypt; New Delhi, India; Jakarta, Indonesia; Cape Town, South Africa; Santiago, Chile; Paris, France; and Cambridge, MA, USA.

As shown in this Thesis (Chapter 3), there has been a relative shift towards research flows to institutions in high-income countries, whereas it was evaluators based in L&MICs who spearheaded the development of impact evaluation approaches in WASH. The majority of the research is therefore being undertaken by consultants and academics based in high income countries, who are at least two steps removed from the realities of WASH programming in L&MICs, and several further away from the lives of the poor. One may reasonably question, therefore, whether incentives are aligned to promote the most poverty-reduction efficient use of development research resources. Funders can influence this by continuing to prioritise applications with capacity building embedded (e.g., PhD studentships in L&MICs). But it seems more could be done to ensure investigators in L&MICs have leading or meaningful roles in WASH evaluation and synthesis research, and that process will be advanced by incentives from funders and publishing bodies.

On the role of plurality in the methods used in impact evaluation and systematic reviews, it is debatable whether analysis of behaviour always requires incorporation of qualitative evidence systematically. Impact evaluations and systematic reviews drawing solely on quantitative evidence from impact evaluations are commonly thought to be unable to answer questions about why interventions are successful or not. However, studies that draw on an explicit theory of change (or logic model) and collect evidence on outcomes along the causal pathway, can explain heterogeneity in findings, even when restricted to quantitative methods only. However, analysis to 'open up the intervention black box' necessarily draws on broader evidence such as from implementation reports and qualitative studies (White, 2018). Alongside the shift to evaluating behaviours in primary studies, it is becoming more common for reviews to incorporate mixed methods. This analysis is highly policy relevant as it enables an understanding of heterogeneity and therefore the circumstances in which review findings are applicable.[168]

Over 30 years ago, Cairncross (1990) noted: "it is striking that there is still no scientific consensus as to whether water supply affects endemic diarrhoeal

[168] For example, the Executive Director of *Banka BioLoo,* an NGO which provides 'sustainable sanitation across India' wrote to the Campbell Collaboration International Development Coordinating Group in appreciation of de Buck et al. (2017) systematic review of sanitation promotional approaches that used mixed-methods, drawing on impact evaluations and qualitative studies.

disease at all, and if it does, whether it achieves this through improvements in water quality, or quantity, or both" (p.311). After reviewing the evidence on WASH impacts, what remains striking is that studies do not typically collect data on distance to the water source, or water consumption (litres per capital per day) and how it is used (e.g., whether consumed or used in bathing). This information is crucial for understanding mechanisms and therefore generalisability of findings. For example, in the review of mortality estimates (Chapter 6, Section 6.6), four studies, only one of which assessed the impact of improved water supply (Hoque et al., 1999), provided information on distance to water supply (Hoque et al., 1999; Emerson et al., 2004; Null et al., 2018; Pickering et al., 2015), and only Pickering et al. (2015) reported water consumption.[169] In addition, some studies appeared to underreport the hygiene component in their discussion of the intervention, including studies of latrines (Pickering et al., 2015; Reese et al., 2019), as well as studies not included due to deworming co-interventions (Miguel and Kremer, 2004). Another study, of handwashing and household water treatment in Pakistan, indicated that participant communities had access to at least two hours of running water per week, but did not report any information on the reliability of the water supply available (Bowen et al., 2012).

Therefore, a final recommendation is for more transparent reporting about the intervention – not just more information about dosage, timing and frequency of community visits, but clear information about the WASH technology itself that is being promoted and the comparison conditions (what WASH technology is available otherwise). For example, where the hygiene messaging is part of the intervention, it should be clearly reported in the study title and abstract.

Thirty-four studies of interventions to improve water supply have been completed in L&MICs (44 studies of water supply alongside sanitation and/or hygiene promotion), of which ten are RCTs (18 of water supply with sanitation or hygiene). Nearly all RCTs measured water supply behaviour but

[169] Hoque et al. (1999) gave the share of households with time to tube well of less than 1 minutes, Emerson et al. (2004) gave the share of households with round-trip to water less than 30 minutes, Pickering et al. (2015) reported share of households within 5 minutes walking time, and Null et al. (2018) reported mean one-way walking time to primary water source. Luby et al. (2018) reported controlling for distance to water source in regression analysis, but did not report the mean distance by intervention group.

only five collected health outcomes data and four measured time use or income. Five more RCTs are ongoing of water supply alone or in combination with hygiene, sanitation and/or weaning foods (Adanu and Wright, 2014; Gertler and Gonzalez-Navarro, 2014; Leder, 2016; Martinez et al., 2017; Morse et al., 2017). With the findings from the evidence reviews in this Thesis, as well as the availability of new studies awaiting review, plus the advent of innovative and cost-effective study designs like natural experiments, consensus on the question posed by Cairncross at the end of the first International Drinking Water Supply and Sanitation Decade may finally be reached early in the new International Decade for Action on Water for Sustainable Development international 2018-28, with the potential to improve the lives of the most disadvantaged people.

References

Abou-Ali, H., El-Azony, H., El-Laithy, H., Haughton, J. and Khandker, S.R. (2009). Evaluating the impact of Egyptian social fund for development programs. World Bank Policy Research Working Paper 4993. Available at: https://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-4993 (accessed 1 September 2020).

Abubakar, I., Aliyu, S.H., Arumugam, C., Usman, N.K. and Hunter, P.R. (2007). Treatment of cryptosporidiosis in immunocompromised individuals: systematic review and meta-analysis. British Journal of Clinical Pharmacology, 63 (4), 387-393. Doi: 10.1111/j.1365-2125.2007.02873.x.

Acey, C., Kisiangani, J., Ronoh, P., Delaire, C., Makena, E., Norman, G., Levine, D., Khush, R. and Peletz, R. (2018). Cross-subsidies for improved sanitation in low income settlements: assessing the willingness to pay of water utility customers in Kenyan cities. World Development, 115, 160-177. https://doi.org/10.1016/j.worlddev.2018.11.006.

Adank, M., Terefe, B., Dickinson, N., Potter, A., Butterworth, J., Mekonta, L., Defere, E. and Bostoen, K. (2017). WASH services in small towns: Midline report for a quasi-randomised control trial to assess impacts of the ONEWASH Plus programme. ONEWASH Plus Programme Report. The Hague, IRC (WASH).

Adanu, R., Hill, A. and Wright, J. (2014). Urban drinking water and health outcomes - early phase study for a randomized controlled trial in Accra, Ghana. AEA RCT Trial Registry. Available at: https://www.socialscienceregistry.org/trials/572.

Aiello, A.E., Coulborn, R.M., Perez, V. and Larson, E.L. (2008). Effect of hand hygiene on infectious disease risk in the community setting: a meta-analysis. American Journal of Public Health, 98 (8), 1372-1381.

Allcott, H. (2015). Site selection bias in program evaluation. The Quarterly Journal of Economics, 130 (3), 1117-1165.

Anand, R., Norrie, J., Bradley, J., McAuley, D. and Clarke, M. (2020). Fool's gold? Why blinded trials are not always best. British Medical Journal, 368 [l6228]. https://doi.org/10.1136/bmj.l6228.

Anderson, E. and Waddington, H. (2007). Aid and the Millennium Development Goal poverty target: how much is required and how should it be allocated. Oxford Development Studies, 35 (1), 1-31.

Angelucci, M. and de Giorgi, G. (2006). Indirect effects of an aid program: the case of PROGRESA and consumption. Discussion Paper No. 1955, January 2006. IZA Institute of Labor Economics, Bonn.

Angrist, J.D. (2004). Treatment effect heterogeneity in theory and practice. The Economic Journal, 114 (March), C52-C83.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91 (434), 444-455.

Angrist, J.D. and Pischke, S. (2009). Mostly harmless econometrics: an empiricist's companion. Princeton University Press, New Jersey.

Anker, M. (1997). The effect of misclassification error on reported cause-specific mortality fractions from verbal autopsy. International Journal of Epidemiology, 26, 1090-6.

Annamalai, T.R., Devkar, G., Mahalingam, A., Benjamin, S., Rajan, S.C. and Deep, A. (2016). What is the evidence on top-down and bottom-up approaches in improving access to water, sanitation, and electricity services in low-income or informal settlements? EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London, London.

Armand, A., Augsburg, B., Bancalari, A. and Trivedi, B. (2020). Community toilet use in Indian slums: willingness-to-pay and the role of informational and supply side constraints. 3ie Impact Evaluation Report 113. International Initiative for Impact Evaluation, New Delhi. https://doi.org/10.23846/DPW1IE113.

Arnold, B.F. and Colford, J.M. Jr. (2007). Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhoea in

developing countries: a systematic review and meta-analysis. American Journal of Tropical Medicine and Hygiene, 76 (2), 354-364.

Arnold, B.F., Khush, R.S., Ramaswamy, P., London, A.G., Rajkumar, P., Ramaprabha, P., Durairaj, N., Hubbard, A.E., Balakrishnan, K. and Colford, J.M. Jr. (2012). Causal inference methods to study nonrandomized, preexisting development interventions. Proceedings of the National Academy of Sciences of the United States of America, 107 (52), 22605–22610.

Arnold, B.F., Khush, R.S., Ramaswamy, P., Rajkumar, P., Durairaj, N., Ramaprabha, P., Balakrishnan, K. and Colford, J.M. Jr. (2015). Short report: reactivity in rapidly collected hygiene and toilet spot check measurements: a cautionary note for longitudinal studies. American Journal of Tropical Medicine and Hygiene, 92 (1), 159-162. Doi:10.4269/ajtmh.14-0306.

Attanasio, O., Kugler, A. and Meghir, C. (2011). Subsidizing vocational training for disadvantaged youth in Colombia: evidence from a randomized trial. American Economic Journal: Applied Economics, 3 (3), 188-220. Doi:10.1257/app.3.3.188.

Aunger, R. and Curtis, V. (2016). Behaviour centred design: towards an applied science of behaviour change. Health Psychology Review, 10 (4), 425-446. http://dx.doi.org/10.1080/17437199.2016.1219673.

Augier, P., Dovis, M. and Lai-Tong, C. (2016). Better access to water, better children's health: a mirage? Oxford Development Studies, 44 (1), 70-92. Doi:10.1080/13600818.2015.1064101.

Austin, C.J. (1993). Water, sanitation, environment and development: chlorinating household water in the Gambia. 19th Water Engineering and Development Centre (WEDC) Conference, Accra, Ghana, 90-92.

Autor, D. (2003). Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing. Journal of Labor Economics, 21 (1), 1-42.

Bacqui, A.H., Black, R.E., Yunus, M.D., Azimul Hoque, A.R. Chowdhury, H.R. and Sack, R.B. (1991). Methodological issues in diarrhoeal diseases

epidemiology: definition of diarrhoeal episodes. International Journal of Epidemiology, 20 (4), 1057-1063. https://doi.org/10.1093/ije/20.4.1057.

Baird, S., Ferreira, F.H.G., Özler, B. and Woolcock, M. (2013). Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: a systematic review, Campbell Systematic Reviews 2013: 8. Doi:10.4073/csr.2013.8.

Baird, S., Hamory Hicks, J., Kremer, M. and Miguel, E. (2016). Worms at work: long-run impacts of a child health investment. Quarterly Journal of Economics, 131 (4), 1637-1680. Doi: 10.1093/qje/qjw022.

Baker, K.K., O'Reilly, C.E., Levine, M.M., Kotloff, K.L., Nataro, J.P., Ayers, T.L., et al. (2016). Sanitation and hygiene-specific risk factors for moderate-to-severe diarrhea in young children in the Global Enteric Multicenter Study, 2007-2011: case-control study. PLoS Medicine, 13 (5), e1002010. Doi:10.1371/journal.pmed.1002010.

Bamberger, M., Rao, V. and Woolcock, M. (2010). Using mixed methods in monitoring and evaluation: experiences from international development, Policy Research Working Paper 5245. The World Bank, Washington, D.C.

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D. and Singh, N. (2012). Can institutions be reformed from within? evidence from a randomized experiment with the Rajasthan police. CEPR Discussion Papers 8869. Doi:10.2139/ssrn.2010854.

Barde, J.A. (2017). What determines access to piped water in rural areas? evidence from small-scale supply systems in rural Brazil. World Development, 95, 88-110.

Barham, T., Macours, K., Maluccio, J.A., Regalia, F., Aguilera, V. and Moncada, M.E. (2014). Assessing long-term impacts of conditional cash transfers on children and young adults in rural Nicaragua. 3ie Impact Evaluation Report 17. International Initiative for Impact Evaluation, New Delhi.

Bärnighausen, T., Røttingen, J.-A., Rockers, P., Shemilt, I. and Tugwell, P. (2017a). Quasi-experimental study designs series – paper 1: history and

introduction. Journal of Clinical Epidemiology, 89, 4-11. Doi:10.1016/j.jclinepi.2017.02.020.

Bärnighausen, T., Tugwell, P., Røttingen, J.-A., Shemilt, I., Rockers, P., Geldsetzer, P., Lavis, J., Grimshaw, J., Daniels, K., Brown, A., Bor, J., Tanner, J., Rashidian, A., Barreto, M., Vollmer, S. and Atun, R. (2017b). Quasi-experimental study designs series – paper 4: uses and value. Journal of Clinical Epidemiology, 89, 21-29. Doi:10.1016/j.jclinepi.2017.03.012.

Bärnighausen, T., Oldenburg , C., Tugwell , P., Bommer , C., Ebert , C., Barreto , M., Djimeu , E., Haber , N., Waddington , H., Rockers , P., Sianesi , B., Bor , J., Fink , G., Valentine , J., Tanner , J., Stanley , T., Sierra , E. Tchetgen , E., Atun , R. and Vollmer, S. (2017c). Quasi-experimental study designs series – paper 7: assessing the assumptions. Journal of Clinical Epidemiology, 89, 53-66. Doi:10.1016/j.jclinepi.2017.02.017.

Barrell, R.A.E., and Rowland, M.G.M. (1980). Commercial milk products and indigenous weaning foods in a rural West African environment: a bacteriological perspective. Journal of Hygiene, 84, 191-202.

Barrera-Osorio, F. and Filmer, D. (2016). Incentivizing schooling for learning evidence on the impact of alternative targeting approaches. Journal of Human Resources, 51 (2), 461-499.

Barrera-Osorio, F., Filmer, D. and McIntyre, J. (2014). An empirical comparison of randomized control trials and regression discontinuity estimations. SREE Conference Abstract, Society for Research on Educational Effectiveness, Evanston, IL.

Barreto, M., Genser, B., Strina, A. Marlucia, Teixeira, M.G., Assis, A.M.O., Rego, R., Teles, C.A., Prado, M.S., Matos, S.M.A., Santos, D.N., dos Santos, L.A. and Cairncross, C. (2007). Effect of city-wide sanitation programme on reduction in rate of childhood diarrhoea in northeast Brazil: assessment by two cohort studies. Lancet, 370, 1622-1628.

Barrientos, A. and Sabates-Wheeler, R. (2011). Strategic complementarities and social transfers: how do PROGRESA payments impact nonbeneficiaries? Applied Economics, 43, 3175-3185.

Barros, A., Ross, D., Fonseca, W., Williams, L. and Moreira-Filho, D. (1999). Preventing acute respiratory infections and diarrhoea in child day care centres. Acta Paediatrica, 88, 1113-18.

Bartram, J. and Cairncross, S. (2010). Hygiene, sanitation, and water: forgotten foundations of health. PLoS Medicine, 7 (11), e1000367.

Batmunkh, O., Chase, C. Galing, E. and La, M.P. (2019) Final evaluation report on integrating sanitation programming in the Pantawid Pamilya Program (Philippines). Impact Evaluation Report. November 12, 2019. The World Bank Group, Washington, D.C.

Beath, A., Christia, F. and Enikolopov, R. (2013). Randomized impact evaluation of Afghanistan's national solidarity program. International Bank for Reconstruction and Development, Washington, D.C.

Begg, C.B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 50 (4), 1088-101. PMID: 7786990.

Behrman, J. and Todd, P. (1999). Randomness in the experimental samples of PROGRESA. International Food Policy Research Institute, Washington, D.C.

Behrman, J., Parker, S. and Todd, P. (2009). Schooling impacts of conditional cash transfers on young children: evidence from Mexico. Economic Development and Cultural Change, 57 (3), 439-78.

Ben Yishay, A., Fraker, A., Guiteras, R., Palloni, G., Shah, N.B., Shirrell, S. and Wang, P. (2017). Microcredit and willingness to pay for environmental quality: evidence from a randomized-controlled trial of finance for sanitation in rural Cambodia. Journal of Environmental Economics and Management, 86, 121-140.

Benova, L., Cumming, O. and Campbell, O.M. (2014). Systematic review and meta-analysis: association between water and sanitation environment and maternal mortality. Tropical Medicine and International Health, 19 (4), 368-87.

Bhatia, R. and Falkenmark, M. (1993). Water resources and the urban poor: innovative approaches and policy imperatives. Working Paper 46877. The World Bank, Washington, D.C.

Bicego, G.T. and Boerma, J.T. (1993). Maternal education and child survival: A comparative study of survey data from 17 countries. Social Science and Medicine, 36, 1207-1227.

Biran, A., Jenkins, M.W., Dabrase, P. and Bhagwat, I. (2011). Patterns and determinants of communal latrine usage in urban poverty pockets in Bhopal, India. Tropical Medicine and International Health, 16 (7), 854-862. Doi:10.1111/j.1365-3156.2011.02764.x.

Biran, A., Schmidt, W.-P., Varadharajan, K.S., Rajaraman, D., Kumar, R., Greenland, K., Gopalan, B., Aunger, R. and Curtis, V. (2014). Effect of a behaviour-change intervention on handwashing with soap in India (SuperAmma): a cluster-randomised trial. Lancet Global Health, 2, pp.e145–154.

Birmingham, M.E., Lee, L.A., Ntakibirora, M., Bizimana, F. and Deming, M.S. (1997). A household survey of dysentery in Burundi: implications for the current pandemic in sub-Saharan Africa. Bulletin of the World Health Organization, 75, 45-53.

Black, R., Cousens, S., Johnson, H.L., Lawn, J.E., Rudan, I., Bassani, D.G., Jha, P., Campbell, H., Walker, C.F., Cibulskis, R., Eisele, T., Liu, L. and Mathers, C. (2010). Global, regional, and national causes of child mortality in 2008: a systematic analysis. Lancet, 375; 9730, 1969-1987.

Black, R.E., Dykes, A.C., Anderson, K.E., Wells, J.G., Sinclair, S.P., Gary, G.W., Hatch, M.H. and Gangarosa, E.J. (1981). Handwashing to prevent diarrhea in day-care centers. American Journal of Epidemiology, 113 (4), 445-451.

Bloom, H.S. (2006). Core analytics of randomised experiments for social research. MDRC Working Papers for Research Methodology, August 2006, MDRC, New York, NY. Available at: https://www.mdrc.org/publication/core-analytics-randomized-experiments-social-research (accessed 17 March 2020).

Bloom, H.S., Michalopoulos, C., Hill, C.J. and Lei, Y. (2002). Can nonexperimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? MDRC Working Papers on Research Methodology, June 2002, MDRC, NY.

Bloomfield, S.F., Rook, G.A., Scott, E.A., Shanahan, F., Stanwell-Smith, R. and Turner, P. (2016). Time to abandon the hygiene hypothesis: new perspectives on allergic disease, the human microbiome, infectious disease prevention and the role of targeted hygiene. Perspectives in Public Health, 136 (4), 213-24. Doi:10.1177/1757913916650225.

Blum, D. and Feachem, R.G. (1983). Measuring the impact of water supply and sanitation investments on diarrheal diseases: problems of methodology. International Journal of Epidemiology, 12 (3), 357-365.

Boisson, S., Kiyombo, M., Sthreshly, L., Tumba, S., Makambo, J. and Clasen, T. (2010). Field assessment of a novel household-based water filtration device: a randomised, placebo-controlled trial in the Democratic Republic of Congo. PLoS ONE, 5 (9), 1-10.

Boisson, S., Schmidt, W.-P., Berhanu, T., Gezahegn, H. and Clasen, T. (2009). Randomized controlled trial in rural Ethiopia to assess a portable water treatment device. Environmental Science and Technology, 43 (15), 5934-5939.

Boisson, S., Stevenson, M., Shapiro, L., Kumar, V., Singh, L.K., Ward, D. and Clasen, T. (2013). Effect of household-based drinking water chlorination on diarrhea among children under five in Orissa, India: a double-blind randomised placebo-controlled trial. PLoS Medicine / Public Library of Science, 10, e1001497.

Bor, J., Moscoe, E., Mutevedzi, P., Newell. M.L. and Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: causal inference without randomised trials. Epidemiology, 25 (5), 729-37.

Borenstein, M. Hedges, L.V., Higgins, J., and Rothstein, H. (2009). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1, 97-111. Doi:10.1002/jrsm.12.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H. (2009). Introduction to meta-analysis. John Wiley and Sons, Chichester.

Borenstein, M., Higgins, J., Hedges, L.V., and Rothstein, H. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Research Synthesis Methods, 8 (1), 5-18. https://doi.org/10.1002/jrsm.1230.

Bosch, C., Hommann, K., Rubio, G., Sadoff, C. and Travers, L. (2002). Water and Sanitation. Chapter 23 in: Klugman, J. (ed.). A sourcebook for poverty reduction strategies. Volume 2 macroeconomic and sectoral approaches. The World Bank, Washington, D.C.

Bose, R. (2010). A checklist for the reporting of randomized control trials of social and economic policy interventions in developing countries: CEDE Version 1.0. Working paper 6. International Initiative for Impact Evaluation, New Delhi.

Bound, J., Jaeger, D. A. and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association, 90, 443-450.

Bowen, A., Agboatwalla, M., Luby, S., Tobery, T., Ayers, T. and Hoekstra, R.M. (2012). Association between intensive handwashing promotion and child development in Karachi, Pakistan. Archives of Paediatric and Adolescent Medicine, 166 (11), 1037-1044.

Bracht, G. and Glass, G. (1968). The external validity of experiments. American Educational Research Journal, 5 (4), 437-474.

Bradford-Hill, A. (1965). The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295-300.

Briceño, B., Coville, A. and Martinez, S. (2015). Promoting handwashing and sanitation: evidence from a large-scale randomized trial in rural Tanzania. Policy Research Working Paper 7164. The World Bank, Washington, D.C.

Briscoe, J., Feachem, R.G. and Rahaman, M.M. (1985). Measuring the impact of water supply and sanitation facilities on diarrhoea morbidity: prospects for case-control methods. WHO/CWS/85.3. CDD/OPR/85.1. World Health Organization, Geneva.

Briscoe, J., Feachem, R.G. and Rahaman, M.M. (1986). Evaluating health impact: water supply, sanitation, and hygiene education. IDRC-248e. UNICEF, International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B) and International Development Research Centre (IDRC), Canada.

Brockerhoff, M. (1990) Rural-to-urban migration and child survival in Senegal. Demography, 27, 601-616.

Brockerhoff, M. and Derose, L.F. (1996). Child survival in East Africa: the impact of preventive health care. World Development, 24, 1841-1857.

Brody, C., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P. and Dworkin, S. (2015). Economic self-help group programs for improving women's empowerment: a systematic review. Campbell Systematic Reviews, 2015:19. Doi:10.4073/csr.2015.19.

Brown, J., Jeandron, A., Cavill, S. and Cumming, O. (2012). Evidence review and research priorities: water, sanitation, and hygiene for emergency response. Department for International Development (DFID), London.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: randomization in practice in development field experiments. American Economic Journal: Applied Economics, 1 (4), 200-232.

Brumback, B. A., He, Z., Prasad, M., Freeman, M. C. and Rheingans, R. (2014). Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial. Statistics in Medicine, 33, 1490-1502. Doi: 10.1002/sim.6049.

Buddelmeyer, H. and Skoufias, E. (2004). An evaluation of the performance of regression discontinuity design on PROGRESA. World Bank Policy Research Working Paper 3386. The World Bank, Washington, D.C.

Butz, W.P., Habicht, J.P. and DaVanzo, J. (1984). Environmental factors in the relationship between breastfeeding and Infant mortality: the role of sanitation and water in Malaysia. American Journal of Epidemiology, 119 (4), 516-25.

Cairncross, S. (1989). Water supply and sanitation: an agenda for research. Journal of Tropical Medicine and Hygiene, 92, 301-314.

Cairncross, S. (1992). Control of enteric pathogens in developing countries. Chapter 7 in: Mitchell, R. (ed). Environmental microbiology. John Wiley and Sons, New York.

Cairncross, S. (2004). The case for marketing sanitation. Sanitation and Hygiene Series Field Note. Water and Sanitation Programme-Africa. The World Bank, Washington, D.C.

Cairncross, S. and Cliff, J.L. (1987). Water use and health in Mueda, Mozambique. Transactions of the Royal Society of Tropical Medicine and Hygiene, 81, 51-54.

Cairncross, S. and Feachem, R.G. (1993). Environmental health engineering in the tropics: an introductory text. Second Edition. John Wiley and Sons, Chichester.

Cairncross, S. and Feachem, R.G. (2018). Environmental health engineering in the tropics: water, sanitation and disease control. Third Edition. Routledge, London.

Cairncross, S. and Kinnear, J. (1992). Elasticity of demand for water in Khartoum, Sudan. Social Science and Medicine, 34 (2), 183-189.

Cairncross, S. and Kolsky, P. (1997). Re: 'water, waste, and well-being: a multicountry study'. American Journal of Epidemiology, 146 (4), 359-360.

Cairncross, S. and Valdmanis, V. (2006). Water supply, sanitation and hygiene promotion. In: Jamison, D. Breman, J., Measham, A., Alleyne, G., Claeson, M., Evans, D., Jha, P., Mills, A. and Musgrove, P. (eds). Disease control priorities in developing countries. 2nd Edition. Oxford University Press, New York.

Cairncross, S., Blumenthal, U., Kolsky, P., Moraes, L, and Tayeh, A. (1996). The public and domestic domains in the transmission of disease. Tropical Medicine and International Health, 1 (1), 27-34.

Cairncross, S., Braide, E. and Bugri, S. (1996). Community participation in the eradication of Guinea worm disease. Acta Tropica, 61, 121-136.

Cairncross, S., Carruthers, I., Curtis, D., Feachem, R., Bradley, D. and Baldwin, G. (1980). Evaluation for village water supply planning. John Wiley and Sons, Chichester.

Cairncross, S., Cumming, O., Schechtman, L., Velleman, Y. and Waddington, H. (2014). Health impacts of sanitation and hygiene. In: Cross, P. and Coombes, Y. (eds). Sanitation and hygiene in Africa: where do we stand?: analysis from the AfricaSan Conference, Kigali, Rwanda. IWA Publishing, London. ISBN 9781780405414 (paperback) 9781780405421 (ebook).

Cairncross, S., Hunt, C., Boisson, S., Bostoen, K., Curtis, V., Fung, I. and Schmidt, W.-P. (2010). Water, sanitation and hygiene for the prevention of diarrhoea. International Journal of Epidemiology, 39, i193-i205. Doi:10.1093/ije/dyq035.

Cameron, L., Shah, M. and Olivia, S. (2013). Impact evaluation of a large-scale rural sanitation project in Indonesia. Policy Research Working Paper 6360, Impact Evaluation Series No. 83. The World Bank, Washington, D.C.

Campbell, D.T. (1984). Can we be scientific in applied social science? Evaluation Studies Review Annual, 9, 26-48.

Campbell, O., Benova, L., Gon, G., Afsana, K. and Cumming, O. (2014). Getting the basic rights – the role of water, sanitation and hygiene in maternal and reproductive health: a conceptual framework. Tropical Medicine and International Health, 20 (3), 252-267.

Capuno, J.J., Tan, C.A. and Fabella, V.M. (2011). Do piped water and flush toilets prevent child diarrhea in rural Philippines? Asia Pacific Journal of Public Health, 2011 Dec 20. PMID:22186402.

Carr-Hill, R., Rolleston, C., Schendel, R. and Waddington, H. (2018). The effectiveness of school-based decision making in improving educational outcomes: a systematic review. Journal of Development Effectiveness, 10 (1), 95-120. Also published as Carr-Hill, R. et al. (2018). The effects of school-based decision making on educational outcomes in low- and middle-income contexts: a systematic review. Campbell Systematic Reviews, 2016:9. Doi:10.4073/csr.2016.9.

Caruso, B.A., Clasen, T.F., Hadley, C., Yount, K.M., Haardörfer, R., Rout, M., Dasmohapatra, M. and Cooper, H.L. (2017). Understanding and defining sanitation insecurity: women's gendered experiences of urination, defecation and menstruation in rural Odisha, India. BMJ Glob Health, 2 (4), e000414. Doi:10.1136/bmjgh-2017-000414.

Casterline, J.B., Cooksey, E.C. and Ismail, A.F.E. (1989). Household income and child survival in Egypt. Demography, 26, 15-35.

Cattaneo, M., Galiani, S. and Gertler, P. (2009). Housing, health and happiness. American Economic Journal: Economic Policy 2009, 1 (1), 75-105.

Chalmers, I. (2014). The art of medicine. The development of fair tests of treatments. Lancet, 383 (9930), 1713-1714.

Chambers, R. (2009). Going to scale with community-led total sanitation: reflections on experience, issues and ways forward. IDS Practice Paper 1, Institute of Development Studies at the University of Sussex, Brighton.

Chambers, R. and von Medeazza, G. (2014). Framing undernutrition: faecally-transmitted infections and the 5 As. IDS Working Paper 2014 No. 450. Institute of Development Studies at the University of Sussex, Brighton.

Chaplin, D., Mamun, A., Protik, A., Schurrer, J., Vohra, D., Bos, K., Burak, H., Meyer, L., Dumitrescu, A., Ksoll, A. and Cook, T. (2017). Grid electricity expansion in Tanzania by Millennium Challenge Corporation (MCC): findings from a rigorous impact evaluation. Mathematica Policy Research (MPR), Princeton, NJ.

Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N. and Morris, R.E. (2018). The internal and external validity of the regression discontinuity design: a meta-analysis of 15 within-study comparisons. Journal of Policy Analysis and Management, 37 (2), 403-429. https://doi.org/10.1002/pam.22051.

Charmarbagwala, R., Ranger, M., Waddington, H. and White, H. (2004). The determinants of child health and nutrition: a meta-analysis. Mimeo, The World Bank, Washington, D.C.

Chauhan, K., Schmidt, W.-P., Aunger, R., Gopalan, B., Saxena, D., Yashobant, S., Patwardhan, V., Bhavsar, P., Mavalankar, D. and Curtis, V. (2020). The 5 Star Toilet Campaign: improving toilet use in rural Gujarat. 3ie Impact Evaluation Report 105. International Initiative for Impact Evaluation, New Delhi. https://doi.org/10.23846/TW14IE105.

Chavasse, D.C., Shier, R.P., Murphy, O.A., Huttly, S.R., Cousens, S.N. and Akhtar, T. (1999). Impact of fly control on childhood diarrhoea in Pakistan: community-randomised trial. Lancet, 353 (9146), 22-25.

Chen, Y., van Geen, A., Graziano, J.H., Pfaff, A., Madajewicz, M., Parvez, F., Iftekhar Hussein, A.Z.M., Slavkovich, V., Islam, T. and Ahsan, H. (2007). Reduction in urinary arsenic levels in response to arsenic mitigation efforts in Araihazar, Bangladesh. Environmental Health Perspectives, 115, 917-923.

Chiba, Y. (2010). Bias analysis of the instrumental variable estimator as an estimator of the average causal effect, Contemporary Clinical Trials, 31, 12-17.

Chief Evaluation Office. (Undated). CEO Regression discontinuity design (RDD) checklist. U.S. Department of Labor, Washington, D.C. Available at: https://www.dol.gov/asp/evaluation/resources/ceo_regression_discontinuity_design_checklist.pdf (accessed 15 December 2018).

Chiller, T.M., Mendoza, C.E., Lopez, M.B., Alvarez, M., Hoekstra, R.M., Keswick, B.H. and Luby, S.P. (2006). Reducing diarrhoea in Guatemalan children: randomized controlled trial of flocculant-disinfectant for drinking-water. Bulletin of the World Health Organization, 84, 28-35.

Chinen, M., de Hoop, T., Alcázar, L., Balarin, M. and Sennett, J. (2017). Vocational and business training to improve women's labour market outcomes in low- and middle-income countries: a systematic review. Campbell Systematic Reviews, 2017:16. Doi:10.4073/csr.2017.16

Chirgwin, H., Cairncross, S., Zehra, D. and Waddington, H. (2021). The effectiveness of interventions promoting uptake of water, sanitation, and hygiene (WASH) in low- and middle-income countries: evidence map. Campbell Evidence and Gap Map.

Cintina, I. and Love, I. (2014). The miracle of microfinance revisited: evidence from propensity score matching. UHERO Working Paper No.

2014-14, The Economic Research Organization at the University of Hawaii, Honolulu, HI.

Clarke, N.E., Clements, A.C.A., Amaral, S., Richardson, A., McCarthy, J.S., McGown J., Bryan, S., Gray, D.J. and Nery, S.V. (2018). (S)WASH-D for Worms: A pilot study investigating the differential impact of school-versus community-based integrated control programs for soil-transmitted helminths. PLoS Neglected Tropical Diseases, 12 (5), e0006389. https://doi.org/10.1371/journal.pntd.0006389.

Clasen, T.F. (2013). Comments from Dr. Thomas Clasen on a draft of the GiveWell Water Quality Report, November 15, 2013. Available at: https://files.givewell.org/files/DWDA%202009/Interventions/Water/Water%20Purification%20Assessment/thomas-clasen-comments.pdf (accessed 1 January 2019).

Clasen, T.F., Alexander, K.T., Sinclair, D., Boisson, S., Peletz, R., Chang, H.H., Majorin, F. and Cairncross, S. (2015). Interventions to improve water quality for preventing diarrhoea. Cochrane Database of Systematic Reviews 2015 (10), CD004794. Doi:10.1002/14651858.CD004794.pub3.

Clasen, T.F., Bostoen, K., Schmidt, W.-P., Boisson, S., Fung, I.C.H., Jenkins, M. W., Scott, B., Sugden, S. and Cairncross, S. (2010). Interventions to improve the disposal of human excreta for preventing diarrhoea (review). Cochrane Database of Systematic Reviews. Cochrane Database of Systematic Reviews, 2010 (6), CD007180. Doi:10.1002/14651858.CD007180.pub2.

Clasen, T.F., Roberts, I., Rabie, T., Schmidt, W.-P. and Cairncross, S. (2006). Interventions to improve water quality for preventing diarrhoea (review). Cochrane Database of Systematic Reviews, 2006 (3), CD004794. Doi:10.1002/14651858.CD004794.pub2.

Coady, D., Grosh, M. and Hoddinott, J. (2003). The targeting of transfers in developing countries: review of experience and lessons. The World Bank, Washington, D.C.

Cocciolo, S., Ghisolfi, S., Habib, A., Rashid, S.M.A. and Tompsett, A. (2020). Access to safe drinking water: experimental evidence from new water sources in Bangladesh. 3ie Impact Evaluation Report 109. International

Initiative for Impact Evaluation, New Delhi. https://doi.org/10.23846/DPW1IE109.

Cochrane Effective Practice and Organisation of Care Group (EPOC). (Undated). Suggested risk of bias criteria for EPOC reviews. Mimeo.

Coffey, D. and Spears, D. (2018). Implications of WASH benefits trials for water and sanitation. Lancet Global Health, 6 (6), PE615. https://doi.org/10.1016/S2214-109X(18)30225-0.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1), 37-46. Doi:10.1177/001316446002000104.

Cole, E.C., Hawkley, M., Rubino, J.R., Crookston, B.T., McCue, K., Dixon, J., Maqelana, T., Cwayi, J., Adams, C. and Kim, J. (2012). Comprehensive family hygiene promotion in peri-urban Cape Town: Gastrointestinal and respiratory illness and skin infection reduction in children aged under 5. South African Journal of Child Health, 6 (4), 109-117. Doi:10.7196/SAJCH.459.

Concato, J., Shah, N. and Horwitz, R.I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England journal of medicine, 342 (25), 1887-92.

Conroy, R.M., Elmore-Meegan, M., Joyce, T., McGuigan, K. and Barnes, J. (1996). Solar disinfection of drinking water and diarrhoea in Maasai children: a controlled field trial. Lancet, 348, 1695-97.

Conroy, R.M., Elmore-Meegan, M., Joyce, T., McGuigan, K. and Barnes, J. (1999). Solar disinfection of water reduces diarrhoeal disease: an update. Archives of Disease in Childhood, 81, 337-338.

Cook, T.D. (2014). Testing causal hypotheses using longitudinal survey data: a modest proposal for modest improvement. National academy of education workshop to examine current and potential uses of NCES longitudinal surveys by the education research community. Available at: https://naeducation.org/wp-content/uploads/2016/10/thomas-cook-nces-longitudinal-surveys.pdf (accessed 10 December 2020).

Cook, T.D., Shadish, W. and Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. Journal of Policy Analysis and Management, 27 (4), 724-750.

Cook, T.D. and Wong, V. (2008). Empirical tests of the validity of the regression discontinuity design: implications for its theory and use in research practice. Annals of Economics and Statistics, GENES (91-92), 127-150.

Cooper, H.M. and Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87 (3), 442-429.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Education Research, 52 (2), 291-302. Doi:10.3102/00346543052002291.

Cowley, D.E. (1995). Protheses for primary total hip replacement: a critical appraisal of the literature, International Journal of Technology Assessment in Health Care, 11 (4), 770-778.

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M. and Thompson, S. (2011). Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence. Medical Research Council, London.

Cronbach, L.J. with editorial assistance by Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64 (3), 391-418. Doi:10.1177/0013164404266386.

Crump, J.A., Okoth, G.O., Slutsker, L., Ogaja, D.O., Keswick, B.H. and Luby, S.P. (2004). Effect of point-of-use disinfection, flocculation and combined flocculation-disinfection on drinking water quality in western Kenya. Journal of Applied Microbiology, 97, 225–231.

Crump, J.A., Otieno, P.O., Slutsker, L., Keswick, B.H., Rosen, D.H., Hoekstra, R.M., Vulule, J.M. and Luby, S.P. (2005). Household based treatment of drinking water with flocculant-disinfectant for preventing

diarrhoea in areas with turbid source water in rural western Kenya: cluster randomised controlled trial. British Medical Journal, 331 (478), 1-6. Doi:10.1136/bmj.38512.618681.EO.

Cumming, O., Arnold, B.F., Ban, R. et al. (2019). The implications of three major new trials for the effect of water, sanitation and hygiene on childhood diarrhea and stunting: a consensus statement. BMC Medicine, 17, 173. https://doi.org/10.1186/s12916-019-1410-x.

Cumming, O., Elliott, M., Overbo, A. and Bartram, J. (2014). Does Global Progress on Sanitation Really Lag behind Water? An Analysis of Global Progress on Community- and Household-Level Access to Safe Water and Sanitation. PLoS ONE 9(12): e114699. https://doi.org/10.1371/journal.pone.0114699.

Curtis, V. (2001). Hygiene: how myths, monsters, and mothers-in-law can promote behaviour change. Journal of Infection, 43, 75-79. Doi:10.1053/jinf.2001.0862

Curtis, V. and Cairncross, S. (2003). Effect of washing hands with soap on diarrhoea risk in the community: a systematic review. Lancet Infectious Diseases, 3 (5), 275-81.

Curtis, V., Kanki, B., Cousens, S., Biallo, I., Kpozehouen, A., Sangaré, M. and Nikiema, M. (2001). Evidence of behaviour change following a hygiene promotion programme in Burkina Faso. Bulletin of the World Health Organization, 79 (6), 518e527.

Curtis, V., Kanki, B., Mertens, T., Traore, E., Diallo, I., Tall, F. and Cousens, S. (1995). Potties, pits and pipes: explaining hygiene behaviour in Burkina Faso. Social Science and Medicine, 41 (3), 383-393.

Da Vanzo, J. (1988). Infant mortality and socioeconomic development: evidence from Malaysian household data. Demography, 25, 581-595.

Da Vanzo, J. and Habicht, J.P. (1986). Infant mortality decline in Malaysia, 1946-1975: the roles of changes in variables and changes in the structure of relationships. Demography, 23, 143-160.

Da Vanzo, J., Butz, W.P. and Habicht, J.P. (1983) How biological and behavioural influences on mortality in Malaysia vary during the first year of life. Population Studies, 37, 381-402.

Dangour, A.D., Watson, L., Cumming, O., Boisson, S., Che, Y., Velleman, Y., Cavill, S., Allen, E. and Uauy, R. (2013). Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children (Review). Cochrane Database of Systematic Reviews, 2013, Aug 1;(8):CD009382. Doi:10.1002/14651858.CD009382.pub2.

Daniels, D., Cousens, S., Makoae, L. and Feachem, R. (1990a). A case-control study of the impact of improved sanitation on diarrhoea morbidity in Lesotho. Bulletin of the World Health Organisation, 68 (4), 455-463.

Daniels, N.A., Simons, S.L., Rodrigues, A., Gunnlaugsson, G., Forster, T.S., Wells, J.G., Hutwagner, L., Tauxe, R.V. and Mintz, E.D. (1999b). First do no harm: making oral rehydration solution safer in a cholera epidemic. American Journal of Tropical Medicine and Hygiene, 60, 1051-1055.

Dar, O.A. and Khan, M.S. (2011). Millennium development goals and the water target: details, definitions and debate. Tropical Medicine and International Health, 16 (5), 540-544. Doi:10.1111/j.1365-3156.2011.02736.x

Das, J., Hammer, J. and Sánchez-Paramo, C. (2009). Remembrance of things past: the impact of recall periods on reported morbidity and health seeking behavior. Mimeo. The World Bank, Washington, D.C.

De Buck, E., van Remoortel, H., Hannes, K., Govender, T., Naidoo, S., Avau, B., Vandeveegaete, A., Musekiwa, A., Vittoria, L., Cargo, M., Mosler, H.-J., Vandekerckhove, P. and Young, T. (2017). Approaches to promote handwashing and sanitation behaviour change in low- and middle-income countries: a mixed method systematic review. Campbell Systematic Reviews, 2017:7. Doi:10.4073/csr.2017.7.

Deaton, A. (2010). Instruments, randomization, and learning about development. Journal of Economic Literature, 48 (June 2010), 424-455.

Deeks, J., Dinnes, R., D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M. and Altman, D.G. (2003). Evaluating non-randomised intervention studies, Health Technology Assessment, 7 (27), 1-192.

Deke, J., Sama-Miller, E. and Hershey, A. (2015). Addressing attrition bias in randomized controlled trials: considerations for systematic evidence reviews. OPRE REPORT 2015-72, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Service, Washington, D.C.

Delea, M.G., Garn, J.V., Synder, J.S., Linabarger, M., Tesfaye, Y. and Freeman, M.C. (2020). The impact of an enhanced demand side sanitation and hygiene promotion on sustained behaviour change and mental well-being in Ethiopia, 3ie Grantee Final Report. International Initiative for Impact Evaluation, New Delhi.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7, 177-188.

Devkar, G.A., Mahalingam, A., Deep, A. and Thillairajan, A. (2013). Impact of private sector participation on access and quality in provision of electricity, telecom and water services in developing countries: a systematic review. Utilities Policy, 27, 65-81.

Devoto, F., Duflo, E., Dupas, P., Pariente, W. and Pons, V. (2012). Happiness on tap: piped water adoption in urban Morocco. American Economic Journal: Economic Policy, 4 (4), 68-99.

DeWilde, C., Milman, A., Flores, Y., Salmeron, J. and Ray, I. (2008). An integrated method for evaluating community-based safe water programmes and an application in rural Mexico. Health Policy and Planning, 23 (6), 452-464.

Diaz, J.J. and Handa, S. (2005). An assessment of propensity score matching as a nonexperimental impact estimator: estimates from Mexico's PROGRESA Program. Working Paper OVE/WP-04/05, July 22, 2005, Office of Evaluation and Oversight, Inter-American Development Bank, Washington, D.C.

Diaz, J.J. and Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: estimates from Mexico's PROGRESA Program. The Journal of Human Resources, 41 (2), 319-345.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. The Journal of the American Medical Association, 263 (10), 1385-1389. Doi:10.1001/jama.1990.03440100097014.

Dobson, D. and Cook, T. (1980). Avoiding type III error in program evaluation: results from a field experiment. Evaluation and Program Planning, 3, 269-276.

Downs, S. and Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. Journal of Epidemiological Community Health, 52, 377-84. Doi:10.1136/jech.52.6.377.

Dreibelbis, R., O'Reilly, K., Bhat, S., Kulkarni, S., Goel, A., Grover, E., Rao, N. and Cumming, O. (2018). The evaluation of a sanitation intervention on sanitation-related emotional and psychological well-being among women and girls in Bihar. 3ie Grantee Final Report. International Initiative for Impact Evaluation (3ie), New Delhi.

Du Preez, M., Conroy, R.M., Ligondo, S., Hennessy, J., Elmore-Meegan, M., Soita, A. and McGuigan, K.G. (2011). Randomized intervention study of solar disinfection of drinking water in the prevention of dysentery in Kenyan children aged under 5 years. Environmental Science and Technology, 45 (21), 9315-9323.

Duflo, E. and Pande, R. (2007). Dams. The Quarterly Journal of Economics, 122 (2), 601-646.

Duflo, E., Banerjee, A., Finkelstein, A., Katz, L.F., Olken, B.A. and Sautmann, A. (2020). In praise of moderation: suggestions for the scope and use of pre-analysis plan for RCTs in economics. Working Paper 26993. National Bureau of Economic Research, Cambridge, MA.

Duflo, E., Greenstone, M., Guiteras, R. and Clasen, T. (2015). Toilets can work: short and medium run health impacts of addressing complementarities and externalities in water and sanitation. Working Paper

21521, September 2015. National Bureau of Economic Research, Cambridge, MA.

Duflo, E., Kremer, M. and Glennerster, R. (2006). Using randomisation in development economics research: a toolkit. Poverty Action Lab, Cambridge, MA.

Dunning, T. (2010). Design-based inference: beyond the pitfalls of regression analysis? In: Collier, D. and Brady, H. (eds). Rethinking social inquiry: diverse tools, shared standards, 2nd Edition. Rowman and Littlefield, Lantham, MD.

Dunning, T. (2012). Natural experiments in the social sciences: a design-based approach. Cambridge University Press, Cambridge.

Dupas, P., Hoffman, V., Kremer, M. and Zwane, A.P. (2016). Targeting health subsidies through a non-price mechanism: a randomized controlled trial in Kenya. Science, 353 (6302), 889-895. Doi:10.1126/science.aaf6288.

Duvendack, M., Hombrados, J.G., Palmer-Jones, R. and Waddington, H. (2012). Assessing 'what works' in international development: meta-analysis for sophisticated dummies. Journal of Development Effectiveness, 4 (3), 456-471.

Easterbrook, P., Gopalan, R., Berlin, J. and Matthews, D. (1991). Publication bias in clinical research. Lancet, 337 (8746), 867-872. Doi:10.1016/0140-6736(91)90201-y.

Effective Public Health Practice Project (EPHPP). (Undated). Quality Assessment Tool for Quantitative Studies. Mimeo.

Egger, M., Smith, G., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. British Medical Journal, 315, 629. Doi:http://dx.doi.org/10.1136/bmj.315.7109.629.

Eisenstein, E.L., Lobach, D.F., Montgomery, P., Kawamoto, K. and Anstrom, K.J. (2007). Evaluating implementation fidelity in health information technology interventions. AMIA 2007 Symposium Proceedings, 211-215.

Ejemot-Nwadiaro, R.I., Ehiri, J.E., Arikpo, D., Meremikwu, M.M. and Critchley, J.A. (2015). Hand washing promotion for preventing diarrhoea.

Cochrane Database of Systematic Reviews, 2015, Sep 3;9:CD004265. Doi:10.1002/14651858.CD004265.pub3.

Ejere, H.O.D., Alhassan, M. B. and Rabiu, M. (2015). Face washing promotion for preventing active trachoma (Review). Cochrane Database of Systematic Reviews, 2015 (2), Apr 18;4:CD003659. Doi:10.1002/14651858.CD003659.pub3.

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M. and Feder, G. (2008). Internal and external validity of cluster randomised trials: systematic review of recent trials. British Medical Journal, 336 (7649), 876-80. Doi:10.1136/bmj.39517.495764.25.

Eldridge, S., Campbell, M., Campbell, M., Drahota, A., Giraudeau, B., Higgins, J., Reeves, B. and Siegfried, N. (2016). Revised Cochrane risk of bias tool for randomized trials (RoB 2.0) Additional considerations for cluster-randomized trials. Available at: https://www.riskofbias.info/welcome/rob-2-0-tool/archive-rob-2-0-cluster-randomized-trials-2016 (accessed 28 October 2020).

Emerson, P.M., Lindsay, S.W., Alexander, N., Bah, M., Dibba, S.-M., Faal, H.B., Lowe, K.O., McAdam, K.P.W.J., Ratcliffe, A.A., Walraven, G.E.L. and Bailey, R.L. (2004). Role of flies and provision of latrines in trachoma control: cluster-randomised controlled trial. Lancet, 363, 1093-98.

Emerson, P.M., Lindsay, S.W., Walraven, G.E., Faal, H., Bogh, C., Lowe, K. and Bailey, R.L. (1999). Effect of fly control on trachoma and diarrhoea. Lancet, 353 (9162), 1401-3.

Eppig, C., Fincher, C.L. and Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. Proceedings of the Royal Society B Biological Science, 277, 3801-3808.

Erasmus, Y. and Jordaan, S. (2019) Scoping study of impact evaluation capacity in Sub-Saharan Africa. Africa Centre for Evidence, Johannesburg. Available at: https://africacentreforevidence.org/project-outputs-impact-evaluation-capacity/ (accessed 24 October 2020).

Ercumen, A., Arnold, B.F., Kumpel, E., Burt, Z., Ray, I., Nelson, K. and Colford, J.M. Jr. (2015a). Upgrading a piped water supply from intermittent to continuous delivery and association with waterborne illness: a matched

cohort study in urban India. PLoS Med 12 (10), e1001892. Doi:10.1371/journal.pmed.1001892.

Ercumen, A., Naser, A.M., Unicomb, L., Arnold, B.F., Colford, J.M. and Luby, S.P. (2015b). Effects of source-versus household contamination of tubewell water on child diarrhea in rural Bangladesh: a randomized controlled trial. PLoS ONE, 10 (3), e0121907. Doi:10.1371/journal.pone.0121907.

Esrey, S.A. (1987). The effect of improved water supplies and sanitation on child growth and diarrheal rates in Lesotho. A thesis presented to the Faculty of the Graduate School of Cornell University, Ithaca, NY.

Esrey, S.A. (1996). Water, waste, and well-being: a multicountry study. American Journal of Epidemiology, 143, 608-623.

Esrey, S.A., and Habicht, J.P. (1988). Maternal literacy modifies the effect of toilets and piped water on infant survival in Malaysia. American Journal of Epidemiology, 127, 1079-87.

Esrey, S.A., Feachem, R.G. and Hughes, J.M. (1985). Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. Bulletin of the World Health Organization, 63 (4), 757-772.

Esrey, S.A., Potash, J.B., Roberts, L. and Schiff, C. (1991). Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. Bulletin of the World Health Organization, 69 (5), 609-621.

Esteves Mills, J. and Cumming, O. (2016). The impact of water, sanitation and hygiene on key health and social outcomes: review of evidence. Department for International Development (DFID) Evidence paper. Sanitation and Hygiene Applied Research for Equity (SHARE) Consortium, London and UNICEF, New York.

Evans, W.D., Pattanayak, S.K., Young, S., Buszin, J., Rai, S. and Bihm, J.W. (2014). Social marketing of water and sanitation products: A systematic review of peer-reviewed literature. Social Science and Medicine, 110, 18-25.

Eysenck, H.J. (1978). An exercise in mega-silliness. American Psychologist, 33 (5), May 1978, 517. http://dx.doi.org/10.1037/0003-066X.33.5.517.a.

Fan, V.Y.-M. and Mahal, A. (2011). What prevents child diarrhoea? The impacts of water supply, toilets, and hand-washing in rural India. Journal of Development Effectiveness, 3 (3), 340-370.

Feachem, R.G., Burns, E. and Cairncross, S. (eds). (1978). Water health and development: an interdisciplinary evaluation. Tri-Med Books, London.

Fewtrell, L. and Colford, J.M. (2004). Water, Sanitation and Hygiene: Interventions and Diarrhoea: A Systematic Review and Meta-analysis. World Bank, Washington, D.C.

Fewtrell, L., Kaufmann, R.B., Kay, D., Enanoria, W., Haller, L. (2005). Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and metaanalysis. Lancet Infectious Diseases, 5, 42-52.

Fiala, N. and Premand, P. (2017). Social accountability and service delivery: experimental evidence from Uganda. Policy Research Working Paper WPS8449. The World Bank, Washington, D.C.

Filmer, D. and Pritchett, L.H. (2001). Estimating wealth effects without expenditure data – or tears: an application to educational enrollments in states of India. Demography, 38 (1), 115-132.

Fink, G., Günther, I. and Hill, K. (2011). The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986-2007. International Journal of Epidemiology, 40 (5), 1196-1204.

Fitzgerald, J. and Gottschalk, P. and Moffitt, R. (1998). An analysis of sample attrition in panel data: the Michigan Panel Study of Income Dynamics, Journal of Human Resources, 33 (2), 251-299.

Foster, A., Karlan, D. and Miguel, T. (2018). Registered reports: piloting a pre-results review process at the Journal of Development Economics. Development Impact Blog, March 09, 2018. The World Bank, Washington, D.C. Available at: https://blogs.worldbank.org/impactevaluations/registered-reports-

piloting-pre-results-review-process-journal-development-economics (accessed 23 March 2020).

Freeman, M., Greene, L., Dreibelbis, R., Saboori, S., Muga, R., Brumback, B. and Rheingans, R. (2012). Assessing the impact of a school-based water treatment, hygiene and sanitation programme on pupil absence in Nyanza Province, Kenya: a cluster-randomized trial. Tropical Medicine and International Health, 17 (3), 380-91. Doi: 10.1111/j.1365-3156.2011.02927.x.

Freeman, M.C., Garn, J.V., Sclar, G.D., Boisson, S., Medlicott, K., Alexander, K.T., Penakalapati, G., Anderson, D., Mahtani, A.G., Grimes, J.E.T., Rehfuess, E.A. and Clasen, T.F. (2017). The impact of sanitation on infectious disease and nutritional status: a systematic review and meta-analysis. International Journal of Hygiene and Environmental Health, 220, 928-949.

Fretheim, A., Soumerai, S.B., Zhang, F., Oxman, A.D. and Ross-Degnan, D. (2013). Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. Journal of Clinical Epidemiology, 66, 883-887.

Fretheim, A., Zhang, F., Ross-Degnan, D., Oxman, A.D., Cheyne, H., Foy, R., Goodacre, S., Herrin, J., Kerse, N., McKinlay, R.J., Wright, A. and Soumerai, S.B. (2015). A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. Journal of Clinical Epidemiology, 68, 324-333.

Friedrich, M., Balasundaram, T., Muralidharan, A., Raman, V.R. and Mosler, H. (2019). Promoting latrine use in Karnataka, India using the RANAS approach to behaviour change. 3ie Grantee Final Report. International Initiative for Impact Evaluation (3ie), New Delhi.

Fuentes, R., Pfütze, T. and Seck, P. (2006). Does access to water and sanitation affect child survival? A five country analysis. United Nations Development Program (UNDP) Human Development Report. UNDP, New York, NY.

Gaarder, M., Masset, E., Waddington, H., White, H. and Mishra, A. (2011). Invisible treatments: placebo and Hawthorne effects in development

programs. Paper presented at 'Mind the Gap' Impact Evaluation Conference, June 15-17, 2011, Cuernavaca, Mexico.

Galdo, V. and Briceño, B. (2005). An impact evaluation of a potable water and sewerage expansion in Quito: is water enough? Working paper OVE/WP-01/05. Office of Evaluation and Oversight (OVE), Inter-American Development Bank (IADB), Washington, D.C.

Galiani, S. and McEwan, P. (2013). The heterogeneous impact of conditional cash transfers. Journal of Public Economics, 103, 85-96.

Galiani, S., Gertler, P. and Schargrodsky, E. (2005). Water for life: the impact of the privatization of water services on child mortality in Argentina. Journal of Political Economy, 113 (1), 83-120.

Galiani, S., Gonzalez-Rozada, M. and Schargrodsky, E. (2009). Water expansion in shantytowns: Health and savings. Economica, 76, 607-622.

Galiani, S., McEwan, P. and Quistorff, B. (2017). External and internal validity of a geographic quasi-experiment embedded in a cluster-randomized experiment. In: Cattaneo, M.D. and Escanciano, J.C. (eds). Regression discontinuity designs: theory and applications. Advances in Econometrics, 38, 195-236. Emerald Publishing Limited, Bingley, UK.

Garn, J.V., Sclar, G.D., Freeman, M.C., Penakalapati, G., Alexander, K.T., Brooks, P., Rehfuess, E.A., Boisson, S., Medlicott, K.O. and Clasen, T.F. (2017). The impact of sanitation interventions on latrine coverage and latrine use: a systematic review and meta-analysis. International Journal of Hygiene and Environmental Health, 220, 329-340.

Gautam, O.P., Schmidt, W.-P., Cairncross, S., Cavill, S. and Curtis, V. (2017). Trial of a novel intervention to improve multiple food hygiene behaviors in Nepal. American Journal of Tropical Medicine and Hygiene, 96 (6), 1415-1426. Doi:10.4269/ajtmh.16-0526.

GBD 2016 Causes of Death Collaborators. (2017a). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet, 390, 1151-210.

GBD 2016 Causes of Death Collaborators. (2017b). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet, 390, 1211-59.

GBD 2016 Causes of Death Collaborators. (2017c). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet, 390, 1151-210.

Gebre, T., Ayele, B., Zerihun, M., House, J.I., Stoller, N.E., Zhou, Z., Ray, K.J., Gaynor, B.D., Porco, T.C., Emerson, P.M., Lietman, T.M. and Keenan, J.D. (2011). Latrine promotion for trachoma: assessment of mortality from a cluster-randomized trial in Ethiopia. The American Journal of Tropical Medicine and Hygiene, 85 (3), 518-523.

Gebretsadik, S. and Gabreyohannes, E. (2016). Determinants of under-five mortality in high mortality regions of Ethiopia: an analysis of the 2011 Ethiopia Demographic and Health Survey data. International Journal of Population Research, 2016, Article ID 1602761, 1-7. http://dx.doi.org/10.1155/2016/1602761.

Geere, J.-A., Bartram, J., Bates, L., Danquah, L., Evans, B., Fisher, M.B., Groce, N., Majuru, B., Mokoena, M.M., Mukhola, M.S., Nguyen-Viet, H., Duc, P.P., Rhoderick Williams, A., Schmidt, W.-P. and Hunter, P.R. (2018). Carrying water may be a major contributor to disability from musculoskeletal disorders in low income countries: a cross-sectional survey in South Africa, Ghana and Vietnam. Journal of Global Health, 8 (1), 1-14. Doi:10.7189/jogh.08.010406.

Geere, J.-A. and Hunter, P. (2020). The association of water carriage, water supply and sanitation usage with maternal and child health. A combined analysis of 49 Multiple Indicator Cluster Surveys from 41 countries. International Journal of Hygiene and Environmental Health, 223, 238-247.

Genser, B., Strina, A., dos Santos, L.A., Teles, C.A., Prado, M.S., Cairncross, S. and Barreto, M.L. (2008). Impact of a city-wide sanitation intervention in a large urban centre on social, environmental and behavioural

determinants of childhood diarrhoea: analysis of two cohort studies. International Journal of Epidemiology, 37, 831-840.

Gertler, P. and Gonzalez-Navarro, M. (2014). Harvesting rainfall: randomized cistern deployment in northeast Brazil. AEA RCT Registry. Available at: https://www.socialscienceregistry.org/trials/561/history/3167.

Gertler, P., Martinez, S., Premand, P. Rawlings, L. and Vermeersch, C. (2010). Impact evaluation in practice. The World Bank, Washington, D.C.

Gertler, P.J., Martinez, S.W. and Rubio-Codina, M. (2012). Investing cash transfers to raise long-term living standards. American Economic Journal: Applied Economics, 4 (1), 164-192. http://dx.doi.org/10.1257/app.4.1.164.

Geruso, M. and Spears, D. (2018). Neighborhood sanitation and infant mortality. American Economic Journal: Applied Economics, 10 (2), 125-62.

Geruso, M. and Spears, S. (2019). Place and demographic inequality: new evidence on open defecation and the Muslim infant survival advantage in India. Working paper submitted to PAA 2020, September 16, 2019.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G.V. (1977). Integrating findings: the meta-analysis of research. Review of Research in Education, 5, 351-379.

Glazerman, S., Levy, D.M. and Myers, D. (2002). Nonexperimental replications of social experiments: a systematic review. Corporation for the Advancement of Policy Evaluation, Washington, D.C.

Glazerman, S., Levy, D.M. and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science, 589 (1), 63-93.

Glewwe, P. and Olinto, P. (2004). Evaluation of the impact of conditional cash transfers on schooling: an experimental analysis of Honduras' PRAF program. Final report for USAID, January 2004, Washington, D.C.

Glewwe, P., Kremer, M., Moulin, S. and Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. Journal of Development Economics, 74, 251-268.

Goldberger, A. (1972). Selection bias in evaluation of treatment effects: the case of interaction. Unpublished manuscript, Madison, WI.

Gómez, F., Ramos Galvan, R., Frenk, S., Cravioto Muñoz, J., Chávez, R. and Vázquez, J. (1956). Mortality in second and third degree malnutrition. Journal of Tropical Paediatrics, 2 (2), 77-82.

Gøtzsche, P.C. (2000). Why we need a broad perspective on meta-analysis. British Medical Journal, 321, 585-6.

Graham, J.P., Hirai, M. and Kim, S.-S. (2016). An analysis of water collection labor among women and children in 24 sub-Saharan African countries. PLoS ONE, 11 (6), e0155981. Doi:10.1371/journal. pone.0155981.

Granados, C. and Sańchez, F. (2014). Water reforms, decentralization and child mortality in Colombia, 1990-2005. World Development, 53, 68-79.

Gray, D.J., Kurscheid, J.M., Park, M.J., Laksono, B., Wang, D., Clements, A.C.A., Hadisaputro, S., Sadler, R. and Stewart, D.E. (2019). Impact of the "BALatrine" intervention on soil-transmitted helminth infections in Central Java, Indonesia: a pilot study. Tropical Medicine and Infectious Disease, 4 (141), 1-10. Doi:10.3390/tropicalmed4040141.

Green, L.W. and Glasgow, R.E. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Evaluation and the Health Professions, 29 (1), 126-153. Doi:10.1177/0163278705284445.

Greene, W.H. (2002). Econometric analysis. Fifth Edition. Prentice Hall, New Jersey.

Greenhalgh, T. (2014). How to read a paper: the basics of evidence-based medicine. Fifth Edition. John Wiley and Sons, Chichester.

Greenland, S. (2000). Principles of multilevel modelling. International Journal of Epidemiology, 29, 158-167.

Gross, R., Schell, B., Molina, M.C., Leao, M.A. and Strack, U. (1989). The impact of improvement of water supply and sanitation facilities on diarrhea and intestinal parasites: A Brazilian experience with children in two low-income urban communities. *Revue saúde publica*, 23 (3), 214-220.

Guiteras, R., Jannat, K., Levine, D.I. and Polley, T. (2015a). Testing disgust- and shame-based safe water and handwashing promotion in urban Dhaka, Bangladesh. 3ie Grantee Final Report. International Initiative for Impact Evaluation (3ie), New Delhi.

Guiteras, R., Levinsohn, J., and Mobarak, A.M. (2015b). Encouraging sanitation investment in the developing world: a cluster-randomized trial. Science Express, 348 (6237), 903-906.

Gundry, S., Wright, J. and Conroy, R. (2004). A systematic review of the health outcomes related to household water quality in developing countries. Journal of Water and Health, 2 (1), 1-13.

Guyatt, G.H., Oxman, A.D., Akl, E.A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Murad, Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P. and Schünemann, H.J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology, 64 (4), 383-394.

Gyimah, S.O. (2002). Ethnicity and infant mortality in sub-Saharan Africa: the case of Ghana. Discussion Paper no. 02-10, November 2002, Population Studies Centre, University of Western Ontario.

Gyorkos, T.W., Maheu-Giroux, M., Blouin, B. and Casapia, M. (2013) Impact of health education on soil-transmitted helminth infections in schoolchildren of the Peruvian Amazon: a cluster-randomized controlled trial. PLoS Neglected Tropical Diseases, 7: e2397. pmid:24069469.

Habicht, J.-P., Victora, C.G. and Vaughan, J.P. (1999). Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. International Journal of Epidemiology, 28, 10-18.

Hahn, J., Todd, P. and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Notes and comments. Econometrica, 69 (1), 201-209.

Han, A.M. and Hlaing, T. (1989). Prevention of diarrhoea and dysentery by hand washing. Transactions of the Royal Society of Tropical Medicine and Hygiene, 83, 128-131.

Hanchett, S., Khan, M.H., Krieger, L. and Kullmann, C. (2011). Sustainability of sanitation in rural Bangladesh. Refereed Paper 1036, 35th WEDC International Conference, Loughborough, UK.

Handa, S. and Maluccio J.A. (2010). Matching the gold standard: comparing experimental and nonexperimental evaluation techniques for a geographically targeted program. Economic Development and Cultural Change, 58 (3), 415-447.

Hansen, B.B. (2008). Covariate balance in simple, stratified and clustered comparative studies, Statistical Science, 23 (2), 219-236.

Hansen, H., Klejnstrup, N.R. and Andersen, O.W. (2013). A comparison of model-based and design-based impact evaluations of interventions in developing countries. American Journal of Evaluation, 34 (3), 320-338.

Hausman, C. and Rapson, D.S. (2018). Regression discontinuity in time: considerations for empirical applications. NBER Working Paper No. 23602. April 2018. National Bureau of Economic Research, Cambridge, MA.

Havelaar, A., Boonyakarnkul, T., Cunliffe, D., Grabow, W., Sobsey, M., Giddings, M., Magara, Y., Ohanian, E., Toft, P., Chorus, I., Cotruvo, J., Howard, G. and Jackson, P. (2003). Guidelines for drinking water quality water borne pathogens, 3. World Health Organization, Geneva.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica, 47 (1), 153-161.

Heckman, J.J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica, 52 (2), 271-320.

Heckman, J.J. and Smith, J.A. (1995). Assessing the case for social experiments. Journal of Economic Perspectives, 9 (2), 85-110.

Heckman, J.J. (1998). Characterizing selection bias using experimental data. Econometrica, 66 (5), 1017-1098.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. The Review of Economic Studies, 64 (4), 605-664.

Hedges, L.V. (1983). A random effects model for effect sizes. Psychological Bulletin, 93 (2), 388-395.

Hedges, L.V. and Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 106-28.

Hedges, L.V., Tipton, E. and Johnson, M.C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1 (1), 39-65. Doi: 10.1002/jrsm.5.

Heijnen, M., Cumming, O., Peletz, R., Ka-Seen Chan, G., Brown, J., Baker, K. and Clasen, T. (2014). Shared sanitation versus individual household latrines: a systematic review of health outcomes. PLoS ONE, 9 (4), e93300. Doi:10.1371/journal.pone.0093300.

Hemming, D., Chirwa, E.W., Dorward, A., Ruffhead, H.J., Hill, R., Osborn, J., Langer, L., Harman, L., Asaoka, H., Coffey, C. and Phillips, D. (2018). Agricultural input subsidies for improving productivity, farm income, consumer welfare and wider growth in low- and lower-middle-income countries. Campbell Systematic Reviews, 2018:4. Doi:https://doi.org/10.4073/csr.2018.4.

Hennegan, J. and Montgomery, P. (2016). Do menstrual hygiene management interventions improve education and psychosocial outcomes for women and girls in low and middle income countries? A systematic review. PLoS ONE, 11 (2), e0146985. https://doi.org/10.1371/journal.pone.0146985

Hernán, M., Hernandez-Diaz, S. and Robins, J.M. (2004). A structural approach to selection bias. Epidemiology, 15 (5), 615-625.

Higgins, J.P.T. and Green, S. (2011). Cochrane handbook for systematic reviews of interventions, Version 5.0.0. London, John Wiley and Sons.

Higgins, J., Deeks, J. and Altman, D. (2011). Special topics in statistics. Chapter 16 in: Higgins J. and Green, S. (eds). Cochrane handbook for systematic reviews of interventions. Version 5.0.1. The Cochrane Collaboration.

Higgins, J.P.T. and Green, S. (eds). (2011). Cochrane handbook for systematic reviews of interventions. Version 5.0.1. The Cochrane Collaboration.

Higgins, J.P.T. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21, 1539-1558. Doi:10.1002/sim.1186.

Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., Savović, J., Schulz, K.F., Weeks, L. and Sterne, J.A.C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. British Medical Journal, 2011;343:d5928.

Higgins, J.P.T., Ramsay, C., Reeves, B.C., Deeks, J.J., Shea, B., Valentine, J.C., Tugwell, P. and Wells, G. (2012). Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. Research Synthesis Methods, 4 (1), 12-25.

Higgins, J.P.T., Sterne, J.A.C., Savović, J., Page, M.J., Hróbjartsson, A., Boutron, I., Reeves, B. and Eldridge, S. (2016). A revised tool for assessing risk of bias in randomized trials. In: Chandler, J., McKenzie, J., Boutron, I. and Welch, V. (editors). Cochrane Methods. Cochrane Database of Systematic Reviews, 2016 (10) Supplement 1. dx.doi.org/10.1002/14651858.CD201601.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. and Welch, V. (2019). Cochrane Handbook for Systematic Reviews of Interventions. Version 6. Available at https://training.cochrane.org/handbook/archive/v6 (accessed 27 July 2021).

Hoekman, J., Frenken, K., de Zeeuw, D. and Heerspink, H.L. (2012). The geographical distribution of leadership in globalized clinical trials. PLoS ONE, 7 (10), e45984. https://doi.org/10.1371/journal.pone.0045984.

Hombrados, J.G. and Waddington, H. (2012). A tool to assess risk of bias for experiments and quasi-experiments in development research. Mimeo. International Initiative for Impact Evaluation, New Delhi.

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D. and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews, (1). Doi:10.1002/14651858.MR000006.pub3

Hoque, B.A., Mahalanabis, D., Alam, M.J. and Islam, M.S. (1995). Post-defecation handwashing in Bangladesh: practice and efficiency perspectives. Public Health, 109 (1), 15-24. https://doi.org/10.1016/S0033-3506(95)80071-9.

Hoque, B.A., Chakraborty, J., Chowdhury, J.T.A., Chowdhury, U.K., Ali, M., El Arifeen, S. and Sack, R.B. (1999). Effects of environmental factors on child survival in Bangladesh: a case control study. Public Health, 113, 57-64.

Howard, G. Bartram, J., Brocklehurst, C., Colford, J.M., Costa, F., Cunliffe, D., Dreibelbis, R., Spindel Eisenberg, J.N., Evans, B., Girones, R., Hrudey, S., Willetts, J. and Wright, C.Y. (2020). COVID-19: urgent actions, critical reflections and future relevance of 'WaSH': lessons for the current and future pandemics. Water and health, 18 (5), 613-630.

Howlader, A.A. and Bhuiyan, M.U. (1999). Mothers' health-seeking behaviour and infant and child mortality in Bangladesh. Asia-Pacific Population Journal, 14, 59-75.

Humphrey, J.H. (2009). Child undernutrition, tropical enteropathy, toilets, and handwashing. Lancet, 374, 1032-1035.

Humphrey, J.H., Mbuya, M.N.N., Ntozini, R., et al. (2019). Independent and combined effects of improved water, sanitation, and hygiene, and improved complementary feeding, on child stunting and anaemia in rural Zimbabwe: a cluster-randomised trial. Lancet Global Health, 7, e132-47.

Hunter, P.R. (2009). Household water treatment in developing countries: comparing different intervention types using meta-regression. Environmental Science and Technology, 43 (23), 8991-8997.

Hunter, P.R., Risebro, H., Yen, M., Lefebvre, H., Lo, C., Hartemann, P., Longuet, C. and Jaquenoud, F. (2014). Impact of the provision of safe drinking water on school absence rates in Cambodia: a quasi-experimental study. Plos One, 9, pp.e91847.

Hutton, G. (2015). Benefits and costs of the water sanitation and hygiene targets for the post-2015 development agenda post-2015 consensus. Working Paper as of 26 January 2015. Copenhagen Consensus Centre, available at: https://www.copenhagenconsensus.com/sites/default/files/water_sanitation_assessment_-_hutton.pdf (accessed 28 August 2020).

Hutton, G. and Haller, L. (2004). Evaluation of the costs and benefits of water and sanitation improvements at the global level. WHO/SDE/WSH/04.04. World Health Organization, Geneva.

Hutton, G. and Varughese, M.C. (2016). The costs of meeting the 2030 Sustainable Development Goal targets on drinking water, sanitation, and hygiene. Water and Sanitation Program Technical Paper, January 2016. The World Bank, Washington, D.C.

Hutton, G., Haller, L. and Bartram, J. (2006). Economic and health effects of increasing coverage of low-cost water and sanitation interventions. Human Development Report Office Occasional Paper 2006/33. UNDP, New York, NY.

Hutton, G., Haller, L. and Bartram, J. (2007). Global cost-benefit analysis of water supply and sanitation interventions. Journal of Water and Health, 5 (4), 481-502.

Iacus, S.M., King, G. and Porro, G. (2012). Causal inference without balance checking: coarsened exact matching. Political Analysis, 20 (1), 1-24.

Imbens, G.W. and Angrist, J.D. (1994). Identification and estimation of local average treatment effects. Econometrica, 62 (2), 467-475.

Imbens, G.W. and Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47 (1), 5-86.

Institute for Resource Development/Macro International. (1990). Demographic and Health Surveys – Phase II: model 'A' questionnaire with commentary for high contraceptive prevalence countries. DHS-II Basic Documentation Number 1. December 1990. Institute for Resource Development/Macro International, Inc., Columbia, Maryland.

Instituto Apoyo. (2000). *Evaluacion de impacto y sostenibilidad de los proyectos de foncodes*. Instituto Apoyo, Tegucigalpa.

International Initiative for Impact Evaluation (3ie). (Undated). Principles for Impact Evaluation. 3ie, New Delhi.

Israel, D. (2007). Impact of increased access and price on household water use in urban Bolivia. The Journal of Environment Development, 16 (58), 58-83.

Jaciw, A.P. (2016). Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: the methodology. Evaluation Review, 40 (3), 199-240. Doi:10.1177/0193841X16664456.

Jadhav, A., Weitzman, A. and Smith-Greenaway, E. (2016). Household sanitation facilities and women's risk of non-partner sexual violence in India. BMC Public Health, 16, 1139. Doi:10.1186/s12889-016-3797-z.

Jain, S., Sahanoon, O.K., Blanton, E., Schmitz, A., Wannemuehler, K.A., Hoekstra, R.M. and Quick, R.E. (2010). Sodium dichloroisocyanurate tablets for routine treatment of household drinking water in periurban Ghana: a randomized controlled trial. The American Journal of Tropical Medicine and Hygiene, 82(1), 16-22.

Jalan J, and Somanathan E. (2008). The importance of being informed: experimental evidence on demand for environmental quality. Journal of Development Economics, 87, pp.14-28.

Jalan, J. and Ravallion, M. (2003). Does piped water reduce diarrhea for children in rural India? Journal of Econometrics, 112 (1), 153-173.

Jenkins, M. and Sugden, S. (2006). Rethinking sanitation: lessons and innovation for sustainability and success in the new Millennium. Human Development Report Office Occasional Paper 27. UNDP, New York, NY.

Jenkins, M.W. (1999). Sanitation promotion in developing countries: why the latrines of Benin are few and far between. PhD dissertation, 428 pages. UC Davis, CA.

Jimenez, A., Cavill, S. and Cairncross, S. (2014). The Neglect of Hygiene Promotion in developing countries, as shown by the Global Analysis and Assessment of Sanitation and Drinking Water survey. Journal of Water, Sanitation and Hygiene for Development, 4 (2), 240-247.

Jimenez, E., Waddington, H., Goel, N., Prost, A., Pullin, A., White, H., Lahiri, S. and Narain, A. (2018). Mixing and matching: using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. Journal of Development Effectiveness, 10 (4), 400-421.

Jolly, R. (2004). Clean water for all. Chapter 13 in: Black, R. and White, H. (eds). Targeting development: critical perspectives on the Millennium Development Goals. Routledge, London.

Jones-Hughes, T., Peters, J., Whear, R., Cooper, C., Evans, H., Depledge, M. and Pearson, M. (2013). Are interventions to reduce the impact of arsenic contamination of groundwater on human health in developing countries effective? A systematic review. Environmental Evidence, 2 (11), 1-32.

Joshi, A. and Amadi, C. (2013). Impact of water, sanitation, and hygiene interventions on improving health outcomes among school children. Journal of Environmental and Public Health, 2013 (984626), 1-10. Doi:10.1155/2013/98462.

Jüni, P., Witschi, A., Bloch, R. and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. JAMA, 282: 1054-60.

Kanaiaupuni, S.M. and Donato, K.M. (1999). Migradollars and mortality: the effects of migration on infant survival in Mexico. Demography, 36, 339-353.

Kar, K. and Chambers, R. (2008). Handbook on community-led total sanitation. Plan UK, London.

Kariuki, J.G., Magambo, K.J., Njeruh, M.F., Muchiri, E.M., Nzioka, S.M. and Kariuki, S. (2012). Effects of hygiene and sanitation interventions on

reducing diarrhoea prevalence among children in resource constrained communities: case study of Turkana District, Kenya. Journal of Community Health, 37, 1178-1184.

Karlan, D.S. and Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. Journal of Development Economics, 98 (1), 71-75.

Katz, J., Carey, V.J., Zeger, S.L. and Sommer, A. (1993). Estimation of design effects and diarrhea clustering within households and villages. American Journal of Epidemiology, 138, 994-1006.

Kawata, K. (1978). Water and other environmental interventions: the minimum investment concept. The American Journal of Clinical Nutrition, 31 (November 1978), 2114-2123.

Khan, M.S.I., Matin, A., Hassan, M.M. and Qader, M.M.A. (1986). Annotated bibliography on water, sanitation and diarrhoeal diseases: roles and relationships. Specialized Bibliography Series No. 12. International Centre for Diarrhoeal Disease Research, Dhaka.

Khan, M.U. (1982). Interruption of shigellosis by handwashing. Transactions of the Royal Society of Tropical Medicine and Hygiene, 76 (2), 164-168.

Khan, M.U. (1987). Limitation of communal latrines in changing the prevalence of parasites and diarrhoeal attack rate in Dhaka peri-urban slums. Environmental Pollution, 47 (3), 187-194.

Kim, S.Y., Park, J.E., Lee, Y.J., Seo, H.-J., Sheen S.-S., Hahn, S., Jang, B.-H. and Son, H.-J. (2013). Testing a tool for assessing the risk of bias for nonrandomised studies showed moderate reliability and promising validity. Journal of Clinical Epidemiology, 66 (4), 408-14. Doi: 10.1016/j.jclinepi.2012.09.016.

King, G. Murray, C.J.L, Salomon, J.A. and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review, 98, 191-207.

King, G., Gakidou, E., Ravishankar, N., Moore, R.T., Lakin, J., Vargas, M., Téllez-Rojo, M.M., Hernández Ávila, J.E., Hernández Ávila, M. and

Hernández Llamas, H. (2007). A “politically robust” experimental design for public policy evaluation, with application to the Mexican Universal Health Insurance Program. Journal of Policy Analysis and Management, 26 (3), 479-506.

Kinkese, D.M., Hang’ombe, M.B., Touré, O., Kinkese, T. and Kangwa, E. (2018). Contamination of complementary weaning foods for children with Escherichia coli and salmonella species in Lusaka District, Zambia. Journal of Preventive and Rehabilitative Medicine, 1 (1), 19-31. Doi:10.21617/jprm.2018.0101.3.

Kirby, M.A., Nagel, C.L., Rosa, G., Umupfasoni, M.M., Iyakaremye, L., Thomas, E.A. and Clasen, T.F. (2017). Use, microbiological effectiveness and health impact of a household water filter intervention in rural Rwanda - a matched cohort study. International Journal of Hygiene and Environmental Health, 220, 1020-1029.

Kirby, M.A., Nagel, C.L., Rosa, G., Zambrano, L.D., Musafiri, S., Ngirabega, J.d.D, Thomas, E.A. and Clasen, T. (2019). Effects of a large-scale distribution of water filters and natural draft rocket-style cookstoves on diarrhea and acute respiratory infection: a cluster-randomized controlled trial in Western Province, Rwanda. PLoS Med, 16 (6): e1002812. https://doi.org/10.1371/journal.pmed.1002812.

Kirchhoff, L.V., McClelland, K.E., Do Carmo Pinho, M., Araujo, J.G., De Sousa, M.A. and Guerrant, R.L. (1985). Feasibility and efficacy of in-home water chlorination in rural north eastern Brazil. Journal of hygiene, 94 (2), 173-180.

Kishor, S. and Parasuraman, S. (1998). Mother's employment and infant and child mortality in India. National Family Health Survey Subject Reports Number 8. Demographic and Health Surveys, Maryland.

Klasen, S., Lechtenfeld, T., Meie, K. and Rieckmann, J. (2011). Impact evaluation report: water supply and sanitation in provincial towns in Yemen. Discussion Paper No. 102. Georg-August-Universitat, Göttingen.

Kluve, J., Puerto, S., Robalino, D., Romero, J.M., Rother, F., Stöterau, J., Weidenkaff, F. and Witte, M. (2017). Interventions to improve the labour market outcomes of youth: a systematic review of training,

entrepreneurship promotion, employment services, and subsidized employment interventions. Campbell Systematic Reviews, 2017:12. Doi:10.4073/csr.2017.12

Koolwal, G. and van de Walle, D. (2010). Access to water, women's work, and child outcomes. Policy Research Working Paper WPS5302. The World Bank, Washington, D.C.

Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H., Panchalingam, S., Wu, Y., Sow, S.O., Sur, D., Breiman, R.F. et al. (2013). Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. Lancet 2013; 382, 209-22.

Krebs, C.J. (2014). Ecological methodology, 3rd Edition (in preparation). Version 6, 21 December 2016. Available from: https://www.zoology.ubc.ca/~krebs/downloads/krebs_chapter_02_2020.pdf (accessed 1 March 2020).

Kremer, M., Leino, J., Miguel, E. and Zwane, A.P. (2011). Spring cleaning: rural water impacts, valuation, and institutions. The Quarterly Journal of Economics, 126, 145-205.

Kremer, M., Miguel, E., Mullainathan, S. and Null, C. (2009) Making water safe: price, persuasion, peers, promoters, or product design? Working Paper. Mimeo. University of California, Berkeley.

Kumar, E.A. (1970). Bore-hole disposal of excreta of children and diarrhoeal morbidity in a rural community. Environmental Health, 12, 155-159.

Kunz, R. and Oxman, A.D. (1998). The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. British Medical Journal, 317 (7167), 1185-1190. Doi:10.1136/bmj.317.7167.1185

Lalonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. American Economic Review, 76 (4), 604-620.

Lansdown, R., Ledward, A., Hall, A., Issae, W., Yona, E., Matulu, J., Mweta, M., Kihamia, C., Nyandini, U. and Bundy, D. (2002). Schistosomiasis,

helminth infection and health education in Tanzania: achieving behaviour change in primary schools. Health Education Research, 17 (4), 425-433.

Larson, E.L, Lin, S.X., Gomez-Pichardo, C. and Della-Latta P. (2004). Effect of antibacterial home cleaning and handwashing products on infectious disease symptoms: a randomized, double-blind trial. Annals of Internal Medicine, 140, 321–329.

Lavis, J. (2009). How can we support the use of systematic reviews in policymaking? PLOS Medicine, 6 (11): e1000141.

Lawlor, D.A., Davey Smith, G., Mitchell, R. and Ebrahim, S. (2006). Adult blood pressure and climate conditions in infancy: a test of the hypothesis that dehydration in infancy is associated with higher adult blood pressure. American Journal of Epidemiology, 163, 608-614. Doi:10.1093/aje/kwj085.

Lawry, S., Samii, C., Hall, R., Leopold, A., Hornby, D. and Mtero, F. (2014). The impact of land property rights interventions on investment and agricultural productivity in developing countries: a systematic review. Campbell Systematic Reviews, 2014:1. Doi:10.4073/csr.2014.1.

Leach, B. and Waddington, H. (2014). How much evidence is enough for action? Evidence Matters blogpost, available at: https://www.3ieimpact.org/blogs/how-much-evidence-enough-action (accessed 18 February 2020).

Leamer, E.E. (1978). Specification Searches. Wiley, New York.

Leamer, E.E. (1983). Let's take the con out of econometrics. American Economic Review, 73 (1), 31-43.

Leder, K. (2016). Evaluating the impact of hygiene and water interventions on diarrhoeal disease in India. ACTRN12616001286437. Australian New Zealand Clinical Trials Registry. Available at: http://www.anzctr.org.au/ACTRN12616001286437.aspx.

Lee, L.F., Rosenzweig, M.R. and Pitt, M.M. (1997). The effects of improved nutrition, sanitation, and water quality on child health in high-mortality populations. Journal of Econometrics, 77 (1), 209-235.

Levy, J.W., Suntarattiwong, P., Simmerman, J.M., Jarman, R.G., Johnson, K., Olsen, S.J. and Chotpitayasunondh, T. (2013). Increased hand washing

reduces influenza virus surface contamination in Bangkok households, 2009-2010. Influenza and Other Respiratory Viruses, 8 (1), 13–16. Doi:10.1111/irv.12204.

Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H.R., Andrews, K.G. and Aryee, M. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study. Lancet, 380, 2224-2260.

Lincoln, F.C. (1930). Calculating waterfowl abundance on the basis of banding returns. United States Department of Agriculture Circular, 118, 1-4.

Lipsey, M.W. and Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioural treatment: confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M.W. and Wilson, D.B. (2001). Practical meta-analysis. SAGE Publishing, London.

Lipsitch, M., Tchetgen, E.T. and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology, 21 (3), 383-388. Doi:10.1097/EDE.0b013e3181d61eeb.

Littell, J.H., Corcoran, J. and Pillai, V. (2008). Systematic reviews and meta-analysis. Oxford University Press, New York, NY.

Liu, L., Johnson, H.L., Cousens, S., Perin, J., Scott, S., Lawn, J.E., Rudan, I., Campbell, H., Cibulskis, R., Li, M., Mathers, C. and Black, R.E. on behalf of Child Health Epidemiology Reference Group of WHO and UNICEF. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet, 379 (9832), 2151-61.

Lockshin, M. and Sajaia, Z. (2004). Maximum likelihood estimation of endogenous switching regression models. The Stata Journal, 3, 282-289.

Loevinsohn, M., Mehta, L., Cuming, K., Nicol, A., Cumming, O. and Ensink, J.H.J. (2015). The cost of a knowledge silo: a systematic re-review of water,

sanitation and hygiene interventions. Health Policy and Planning, 2015;30:660-674. Doi:10.1093/heapol/czu039.

Lokshin, M. and Yemtsov, R. (2003). Evaluating the impact of infrastructure rehabilitation projects on household welfare in rural Georgia. Policy Research Working Paper 3155. The World Bank, Washington, D.C.

Luby, S.P., Agboatwalla, M., Feikin, D.R., Painter, J., Billhimer, W., Altaf, A. and Hoekstra, R.M. (2005). Effect of handwashing on child health: a randomised controlled trial. Lancet, 366, 225-233.

Luby, S.P., Agboatwalla, M., Hoekstra, R.M., Rahbar, M.H., Billhimer, W. and Keswick, B.H. (2004). Delayed effectiveness of home-based interventions in reducing childhood diarrhea, Karachi, Pakistan. American Journal of Tropical Medicine and Hygiene, 71 (4), 420-427.

Luby, S.P., Agboatwalla, M., Painter, J., Altaf, A., Billhimer, W. and Hoekstra, R.M. (2004). Effect of intensive handwashing promotion on childhood diarrhea in high-risk communities in Pakistan: A randomized controlled trial. JAMA, 291 (21), 2547-2554.

Luby, S.P., Agboatwalla, M., Painter, J., Altaf, A., Billhimer, W. and Keswick, B. (2006). Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Tropical Medicine and International Health, 11 (4), 479–489.

Luby, S.P., Mendoza, C., Keswick, B.H., Chiller, T.M. and Hoekstra, R.M., (2008). Difficulties in bringing point-of-use water treatment to scale in rural Guatemala. American journal of tropical medicine and hygiene, 78 (3), 382-387.

Luby, S.P., Rahman, M., Arnold, B.F., Unicomb, L., Ashraf, S., Winch, P.J., Stewart, C.P., Begum, F., Hussain, F., Benjamin-Chung, J., Leontsini, E., Naser, A.M., Parvez, S.M., Hubbard, A.E., Lin, A., Nizame, F.A., Jannat, K., Ercumen, A., Ram, P.K., Das, K.K., Abedin, J., Clasen, T.F., Dewey, K.G., Fernald, L.C., Null, C., Ahmed, T. and Colford, J.M. Jr. (2018). Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. Lancet Global Health, 6, pp.e302-e315.

Lucas, P.J., Cabral, C. and Colford, J.M. Jr. (2011). Dissemination of drinking water contamination data to consumers: a systematic review of impact on consumer behaviors. PLoS ONE, 6, pp.e21098.

Lule, J.R., Memin, J., Ekwaru, J.P., Malamba, S., Downing, R., Ransom, R., Nakanjako, D., Wafulo, W., Hughes, P., Bunnell, R., Kaharuza, F., Coutinho, A., Kigozi, A. and Quick, R. (2005). Effect of home-based water chlorination and safe storage on diarrhea among persons with human immunodeficiency virus in Uganda. The American Journal of Tropical Medicine and Hygiene, 73 (5), 926-933.

Macassa, G., Ghilagaber, G., Bernhardt, E. and Burstrom, B. (2004). Contribution of household environment factors to urban childhood mortality in Mozambique. East African Medical Journal, 81 (8), 408-414.

Macura, B., del Duca, L., Soto, A., Carrard, N., Gosling, L., Hannes, K., Thomas, J., Sara, L., Sommer, M., Sharma Waddington, H. and Dickin, S. (2021). PROTOCOL: What is the impact of complex WASH interventions on gender and social equality outcomes in low- and middle-income countries? A mixed-method systematic review protocol. Campbell Systematic Reviews, 17 (2), e1164. https://doi.org/10.1002/cl2.1164.

Majorin, F., Torondel, B., Chan, G.K.S. and Clasen, T. (2019). Interventions to improve disposal of child faeces for preventing diarrhoea and soil-transmitted helminth infection. Cochrane Database of Systematic Reviews, 2019 (9), CD011055. Doi:10.1002/14651858.CD011055.pub2.

Makutsa, P., Nzaku, K., Ogutu, P., Barasa, P., Ombeki, S., Mwaki, A. and Quick, R.E. (2001). Challenges in implementing a point-of-use water quality intervention in rural Kenya. American Journal of Public Health, 91 (10), 1571e1573.

Maluccio, J., and Flores, R. (2004). Impact evaluation of a conditional cash transfer program: the Nicaraguan Red de Protección Social. FCND Discussion Paper No. 184, Food Consumption and Nutrition Division, International Food Policy Research Institute, Washington, D.C.

Maluccio, J., and Flores, R. (2005). Impact evaluation of a conditional cash transfer program: the Nicaraguan Red de Protección Social. Research

Report No. 141. International Food Policy Research Institute, Washington, D.C.

Manun'Ebo, M., Cousens, S., Haggerty, P., Kalengaie, M., Ashworth, A. and Kirkwood, B. (1997). Measuring hygiene practices: a comparison of questionnaires with direct observations in rural Zaire. Tropical Medicine and International Health, 2 (11), 1015-1021.

Mara, D. (2017). Review paper: the elimination of open defecation and its adverse health effects: a moral imperative for governments and development professionals. Journal of Water, Sanitation and Hygiene for Development, 7 (1), 1-12. Doi:10.2166/washdev.2017.027.

Mark, M.M. and Lenz-Watson, A.L. (2011). Ethics and the conduct of randomized experiments and quasi-experiments in field settings. In: Printer, A.T. and Sterba, S.K. Handbook of ethics in quantitative methodology. Routledge, Avignon, Oxon.

Martinez, S., Vidal, C. and Sturzenegger, G. (2017). The effects of a comprehensive water and sanitation program in small rural communities in Bolivia. AEA RCT Trial Registry. Available at: https://www.socialscienceregistry.org/trials/2262.

Maslow, A. (1943). A theory of human motivation. Psychological Review, 50, 370-396.

Masset, E. (2019). Impossible generalisations: metaanalyses of education interventions. 19th June 2019, RISE Conference, Washington, D.C. Available at: https://www.riseprogramme.org/sites/www.riseprogramme.org/files/inline-files/Masset_Session5.pdf (accessed 1 April 2020).

Masset, E. (2020). Is there a role for machine learning in the systematic review of evidence? D Case Studies. In: García, O.A. and Kotturi, P. (eds). Information and Communication Technologies for Development Evaluation. Routledge, London.

Masset, E. and White, H. (2003). Infant and child mortality in Andhra Pradesh: analysing changes over time and between states. Young Lives Working Paper No. 18,

Masset, E., Gaarder, M., Beynon, P. and Chapoy, C. (2013). What is the impact of a policy brief? Results of an experiment in research dissemination. Journal of Development Effectiveness, 5 (1), 50-63.

Massey, K. and SHARE. (2011). Insecurity and shame: exploration of the impact of the lack of sanitation on women in the slums of Kampala, Uganda. Cited in WaterAid, undated. Nowhere to go: how a lack of safe toilets threatens to increase violence against women in slums. Available at: www.wateraid.org/se/~/media/Files/Sweden/nowhere-to-go.pdf

Mäusezahl, D., Christen, A., Pacheco, G.D., Tellez, F.A., Iriarte, M., Zapata, M.E., Cevallos, M., Hattendorf, J., Cattaneo, M.D., Arnold, B., Smith, T.A. and Colford, J.M.Jr. (2009). Solar drinking water disinfection (SODIS) to reduce childhood diarrhoea in rural Bolivia: a cluster-randomized, controlled trial. PLoS Medicine / Public Library of Science, 6, e1000125.

Mbakaya, B.C., Lee, P.H. and Lee, R.L.T. (2017). Hand hygiene intervention strategies to reduce diarrhoea and respiratory infections among schoolchildren in developing countries: a systematic review. International Journal of Environmental Research and Public Health, 14 (371), 1-14. Doi:10.3390/ijerph14040371.

McCrary, J. (2006). Manipulation of the running variable in the regression discontinuity design: a density test. Mimeo, University of Michigan.

McGuinness, S.L., O'Toole, J., Forbes, A.B., Boving, T.B., Patil, K., D'Souza, F., Gaonkar, C.A., Giriyan, A., Barker, S.F., Cheng, A.C., Sinclair, M. and Leder, K. (2020). A stepped wedge cluster-randomized trial assessing the impact of a riverbank filtration intervention to improve access to safe water on health in rural India. American Journal of Tropical Medicine and Hygiene, 102 (3), 497-506. Doi:10.4269/ajtmh.19-0260.

McKenzie, D., Stillman, S. and Gibson, J. (2010). How important is selection? Experimental vs nonexperimental measures of the income gains from migration. Journal of the European Economic Association, 8 (4), 913-945.

McSweeney, B.G. (1979). Collection and analysis of data on rural women's time use. Studies in Family Planning, 10 (11/12), 379-383.

Meddings, D.R., Ronald, L.A., Marion, S., Pinera, J.F. and Oppliger, A. (2004). Cost effectiveness of a latrine revision programme in Kabul, Afghanistan. Bulletin of the World Health Organization, 82, 281-289.

Meenakshi, J.V., Banerji, A., Mukherji, A. and Gupta, A. (2013). Impact of metering of agricultural tubewells on groundwater use and informal groundwater irrigation services markets in West Bengal, India. Impact Evaluation Report 4. 3ie, New Delhi.

Mellington, N. and Cameron, L. (1999). Female education and child mortality in Indonesia. Bulletin of Indonesian Economic Studies, 35, 115-44.

Messou, E., Sangaré, S.V., Josseran, R., Le Corre, C. and Guélain, J. (1997). Effect of hygiene and water sanitation and oral rehydration on diarrhoea and mortality of children under five in rural area of Côte d'Ivoire. *Bulletin de la Societe de Pathologie Exotique*, 90 (1), 44-47.

M'Gonigle, G.C.M. and Kirby, J. (1937). Poverty and public health. Victor Gollancz Ltd, London.

Miguel, E. and Kremer, M. (2003). Networks, social learning, and technology adoption: the case of deworming drugs in Kenya. Mimeo. Department of Economics, University of California, Berkeley.

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

Moe, C.L., Sobsey, M.D., Samsa, G.P. and Mesolo, V. (1991). Bacterial indicators of risk of diarrhoeal disease from drinking-water in the Philippines. Bulletin of the World Health Organization, 69 (3), 305-317.

Moher, D. (1998). CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. Consolidated Standards of Reporting Trials. JAMA, 279, 1489-91.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M. and Altman, D.G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. British Medical Journal, 340, 1-28. https://doi.org/10.1136/bmj.c869.

Molina, E., Carella, L., Pacheco, A., Cruces, G. and Gasparini, L. (2016). Community monitoring interventions to curb corruption and increase access and quality of service delivery in low- and middle-income countries. Campbell Systematic Reviews, 2016:8. Doi:10.4073/ csr.2016.8.

Montgomery, P., Hennegan, J., Dolan, C., Wu, M., Steinfield, L. and Scott, L. (2016). Menstruation and the cycle of poverty: a cluster quasi-randomised control trial of sanitary pad and puberty education provision in Uganda. PLoS One, 11, pp.e0166122.

Montgomery, P., Underhill, K., Gardner, F., Operario, D. and Mayo-Wilson, E. (2013). The Oxford Implementation Index: a new tool for incorporating implementation data into systematic reviews and meta-analyses. Journal of Clinical Epidemiology, 66, 874-882. http://dx.doi.org/10.1016/j.jclinepi.2013.03.006.

Moraes, L.R.S., Cancio, J.A. and Cairncross, S. (2004). Impact of drainage and sewerage on intestinal nematode infections in poor urban areas in Salvador, Brazil. Transactions of the Royal Society of Tropical Medicine and Hygiene, 98, 197-204.

Moraes, L.R.S., Cancio, J.A., Cairncross, S. and Huttly, S.R.A. (2003). Impact of drainage and sewerage on diarrhoea in poor urban areas in Salvador, Brazil. Transactions of the Royal Society of Tropical Medicine and Hygiene, 97, 153-158.

Morris, J.F., Murphy, J., Fagerli, K., Schneeberger, C., Jaron, P., Moke, F., Juma, J., Ochieng, J.B., Omore, R., Roellig, D., Xiao, L., Priest, J.W., Narayanan, J., Montgomery, J., Hill, V., Mintz, E., Ayers, T.L. and O'Reilly, C.E. (2018). A randomized controlled trial to assess the impact of ceramic water filters on prevention of diarrhea and cryptosporidiosis in infants and young children - Western Kenya, 2013. The American Journal of Tropical Medicine and Hygiene, 98 (5), 1260-1268.

Morris S.S., Cousens S.N., Kirkwood B.R., Arthur P. and Ross, D.A. (1996). Is prevalence of diarrhea a better predictor of subsequent mortality and weight gain than diarrhea incidence? American journal of epidemiology, 144 (6), 582-8.

Morris, S.S., Black, R.E. and Tomaskovic, L. (2003). Predicting the distribution of under-five deaths by cause in countries without adequate vital registration systems. International Journal of Epidemiology, 32, 1041-1051. Doi:10.1093/ije/dyg241.

Morris, S., Olinto, P., Flores, R., Nilson, E. and Figueiró, A. (2004). Conditional cash transfers are associated with a small reduction in the rate of weight gain of preschool children in Northeast Brazil. Journal of Nutrition, 134 (9), 2336-2341.

Morse, T., Msiska, T. and Cairncross, S. (2017). Evaluating the effect of integration of hygiene of weaning foods with water and sanitation on diarrhoeal disease in under fives (Malawi). Pan African Clinical Trials Registry. Available at: http://www.pactr.org/ATMWeb/appmanager/atm/atmregistry?dar=true&tNo=PACTR201703002084166.

Moscoe, E., Bor, J. and Bärnighausen T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. Journal of Clinical Epidemiology, 68 (2), 122-33. Doi:10.1016/j.jclinepi.2014.06.021.

Mosley, W. and Chen, L. (1984). An analytical framework for the study of child survival in developing countries. Population and Development Review, Supplement to Volume 10, 25-45.

Motarjemi, Y., Käferstein, F., Moy, G. and Quevedo, F. (1993). Contaminated weaning food: a major risk factor for diarrhoea and associated malnutrition. Bulletin of the World Health Organization, 71(1), 79–92.

Mukherjee, C., White, H. and Wuyts, M. (1997). Econometrics and data analysis for developing countries. Priorities for development economics. Routledge, London.

Murthy, G.V., Goswami, A., Narayanan, S. and Amar, S. (1990). Effect of educational intervention on defaecation habits in an Indian urban slum. Journal of Tropical Medicine and Hygiene, 93, 189-193.

National Institute for Health and Clinical Excellence (NICE). (2009). Quality appraisal checklist – quantitative intervention studies. In: Methods

for the development of NICE public health guidance (second edition), April 2009, NICE, London.

Newell, D.J. (1962). Errors in the interpretation of errors in epidemiology. American Journal of Public Health, 52, 1925-1928.

Newhouse, J.P. and McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. Annual Review of Public Health, 19, 17-34.

Nicholson, J.A., Naeeni, M., Hoptroff, M., Matheson, J.R., Roberts, A.J., Taylor, D., Sidibe, M., Weir, A.J., Damle, S.G. and Wright, R.L. (2014). An investigation of the effects of a hand washing intervention on health outcomes and school absence using a randomised trial in Indian urban communities. Tropical Medicine and International Health, 19 (3), 284-292.

Norman, G., Pedley, S. and Takkouche, B. (2010). Effects of sewerage on diarrhoea and enteric infections: a systematic review and meta-analysis. Lancet Infectious Diseases, 10 (8), 536-44. Doi:10.1016/S1473-3099(10)70123-7.

Null, C., Kremer, M., Miguel, E., Hombrados, J.G., Meeks, R. and Zwane, A.P. (2012). Willingness to pay for cleaner water in less developed countries: systematic review of experimental evidence, 3ie Systematic Review 6. International Initiative for Impact Evaluation, London.

Null, C., Stewart, C.P., Pickering, A.J., Dentz, H.N., Arnold, B.F., Arnold, C.D., Benjamin-Chung, J., Clasen, T., Dewey, K.G., Fernald, L.C.H., Hubbard, A.E., Kariger, P., Lin, A., Luby, S.P., Mertens, A., Njenga, S.M., Nyambane, G., Ram, P.K. and Colford, J.M. Jr. (2018). Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial. Lancet Global Health, 6, pp.e316-e329.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Anaiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches, Systematic Reviews, 4 (5), 1-22.

Oberhelman, R.A., Gilman, R.H., Sheen, P., Cordova, J., Zimic, M., Cabrera, L., Meza, R. and Perez, J. (2006). An intervention-control study of corralling of free-ranging chickens to control Campylobacter infections among

children in a Peruvian periurban shantytown. American Journal of Tropical Medicine and Hygiene, 74 (6), 1054-9.

Olusanya, B.O. and Ofovwe, G.E. (2010). Predictors of preterm births and low birthweight in an inner-city hospital in sub-Saharan Africa. Maternal and Child Health Journal, 14 (6), 978-986.

Onyango-Ouma, W., Aagaard, H.J., Jensen, B.B. (2005). The potential of schoolchildren as health change agents in rural western Kenya. Social Science and Medicine, 61 (8), 1711e1722.

Oosterbeek, H., Ponce, J. and Schady, N. (2008). The impact of cash transfers on school enrolment: evidence from Ecuador. Policy Research Working Paper No. 4645. The World Bank, Washington, D.C.

Orgill, J. (2017). Water, sanitation, and development: household preferences and long-term impacts. Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of University Program in Environmental Policy in the Graduate School of Duke University, Raleigh, NC.

Oya, C., Schaefer, F., Skalidou, D., McCosker, C. and Langer, L. (2017). Effects of certification schemes for agricultural production on socio-economic outcomes in low- and middle-income countries: a systematic review. Campbell Systematic Reviews, 2017:3. Doi:10.4073/csr.2017.3

Palmer, T.M. and Sterne, J.A.C. (2016). Meta-analysis in Stata: an updated collection from the Stata journal. Second edition. Stata Press, College Station, TX.

Parikh, P. and McRobie, A. (2009). Engineering as a tool for improving human habitat. International Journal of Management and Decision Making, 10 (3/4), 270-81.

Pattanayak, S.K., Poulos, C., Yang, J.-C. and Patil, S. (2010). How valuable are environmental health interventions? Evaluation of water and sanitation programmes in India. Bulletin of the World Health Organization, 88, 535-42.

Patil, S.R., Arnold, B.F., Salvatore, A.L., Briceño, B., Ganguly, S., Colford, J.M.Jr and Gertler, P.J. (2014). The effect of India's total sanitation

campaign on defecation behaviors and child health in rural Madhya Pradesh: a cluster randomized controlled trial. PLOS Medicine, 11 (8), pp.e1001709.

Pearl, J. and Mackenzie, D. (2018). The book of why: the new science of cause and effect. Penguin Random House, London.

Peletz, R., Kisiangani, J., Ronoh, P., Cock-Esteb, A., Chase, C., Khush, R. and Luoto, J. (2019). Assessing the demand for plastic latrine slabs in rural Kenya. American Journal of Tropical Medicine and Hygiene, 101 (3), 555-565. Doi:10.4269/ajtmh.18-0888.

Peletz, R., Simunyama, M., Sarenje, K., Baisley, K., Filteau, S., Kelly, P. and Clasen, T. (2012) Assessing water filtration and safe storage in households with young children of HIV-positive mothers: a randomized, controlled trial in Zambia. PLoS, ONE 7 (10), e46548. Doi:10.1371/journal.pone.0046548.

Peters, J., Sutton, A., Jones, D., Abrams, K. and Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. Journal of Clinical Epidemiology, 61 (10), 991-996.

Peterson, E.A., Roberts, L., Toole, M.J. and Peterson, D.A. (1998). The effect of soap distribution on diarrhoea: Nyamithuthu refugee camp. International Journal of Epidemiology, 27, 520-24.

Petrosino, A., Morgan, C., Fronius, T.A., Tanner-Smith, E.E. and Boruch, R.F. (2012). Interventions in developing nations for improving primary and secondary school enrollment of children: a systematic review. Campbell Systematic Reviews, 2012:19. Doi:10.4073/csr.2012.19.

Phillips, D., Coffey, C., Tsoli, S., Stevenson, J., Waddington, H., Eyers, J., White, H. and Snilstveit, B. (2017). A map of evidence maps relating to sustainable development in low- and middle-income countries. 3ie Evidence Gap Map Report 10. 3ie, London.

Phillips, D., Waddington. H. and White, H. (2014). Better targeting of farmers as a channel for poverty reduction: a systematic review of Farmer Field Schools targeting. Development Studies Research, 1 (1), 113-136.

Pickering, A.J., Djebbari, H., Lopez, C., Coulibaly, M. and Alzua, M.-L. (2015). Effect of a community-led sanitation intervention on child diarrhoea and child growth in rural Mali: a cluster-randomised controlled trial. Lancet Global Health, 3, pp.e701–11.

Pickering, A.J., Null, C., Winch, P.J., Mangwadu, G., Arnold, B.F., Prendergast, A.J., Njenga, S.M., Rahman, M., Ntozini, R., Benjamin-Chung, J., Stewart, C.P., Huda, T.M.N., Moulton, L.H., Colford, J.M. Jr., Luby, S.P. and Humphrey, J.H. (2019). The WASH Benefits and SHINE trials: interpretation of WASH intervention effects on linear growth and diarrhoea. Lancet Global Health, 2019; 7: e1139-46.

Pirog, M., Buffardi, A., Chrisinger, C., Singh, P. and Briney, J. (2009). Are alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. Journal of Policy Analysis and Management, 28, 169-172. 10.1002/pam.20411.

Piza, C., Cravo, T., Taylor, L., Gonzalez, L., Musse, I., Furtado, I., Sierra, A.C. and Abdelnour, S. (2016). The impact of business support services for small and medium enterprises on firm performance in low- and middle-income countries: a systematic review. Campbell Systematic Reviews, 2016:1. Doi:10.4073/csr.2016.1

Popkin, B.M. and Solon, F.S. (1976). Income, time, the working mother and child nutriture, Journal of Tropical Paediatrics, 22 (4), 156-166.

Porter, G., Hampshire, K., Dunn, C., Hall, R., Levesley, M., Burton, K., Robson, S., Abane, A., Blell, M. and Panther, J. (2013). Health impacts of pedestrian head-loading: a review of the evidence with particular reference to women and children in sub-Saharan Africa. Social Science and Medicine, 88, 90-97.

Poulos, C., Pattanayak, S.K. and Jones K. (2006). A guide to water and sanitation sector impact evaluations (English). Doing impact evaluations series no. 4. The World Bank, Washington, D.C.

Prüss-Üstün, A., Bos, R., Gore, F. and Bartram, J. (2008). Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health. World Health Organization, Geneva.

Prüss-Üstün, A., Mathers, C., Corvalán, C. and Woodward, A. (2003). 3 The global burden of disease concept. Introduction and methods: assessing the environmental burden of disease at national and local levels. Environmental burden of disease. World Health Organization, Geneva.

Prüss-Ustün, A., Wolf, J., Bartram, J., Clasen, T., Cumming, O., Freeman, M.C., Gordon, B., Hunter, P.R., Medlicott, K. and Johnston, R. (2019). Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. International Journal of Hygiene and Environmental Health, 222, 765-777. https://doi.org/10.1016/j.ijheh.2019.05.004.

Quick, R., Kimura, A., Thevos, A., Tembo, M., Shamputa, I., Hutwagner, L. and Mintz, E. (2002). Diarrhea prevention through household-level water disinfection and safe storage in Zambia. American Journal of Tropical Medicine and Hygiene 66 (5), 584-589.

Rabbani, A. (2017). Can leaders promote better health behavior? Learning from a sanitation and hygiene communication experiment in rural Bangladesh. South Asian Network for Development and Environmental Economics (SANDEE) Working Paper No. 122-17, Kathmandu.

Rabie, T. and Curtis, V. (2006). Handwashing and risk of respiratory infections: a quantitative systematic review. Tropical Medicine and International Health, 11 (3), 258-67.

Rabiu, M., Alhassan, M.B., Ejere, H.O.D. and Evans, J.R. (2012). Environmental sanitary interventions for preventing active trachoma. Cochrane Database of Systematic Reviews, 2012 (2), CD004003. Doi:10.1002/14651858.CD004003.pub4.

Rader, T., Pardo, J.P., Stacey, D., Ghogomu, E., Maxwell, L.J., Welch, V.A., Singh, J.A., Buchbinder, R., Légaré, F., Santesso, N., Toupin April, K., O'Connor, A.M., Wells, G.A., Winzenberg, T.M., Johnston, R., Tugwell, P. and the Cochrane Musculoskeletal Group Editors. (2013). Update of strategies to translate evidence from Cochrane musculoskeletal group systematic reviews for use by various audiences. Journal of Rheumatology, 41 (2), 206-215. Doi:10.3899/jrheum.121307.

Ram, P.K., Nasreen, S., Kamm, K., Allen, J., Kumar, S., Rahman, M.A., Zaman, K., El Arifeen, S. and Luby, S.P. (2017). Impact of an intensive perinatal handwashing promotion intervention on maternal handwashing behavior in the neonatal period: findings from a randomized controlled trial in rural Bangladesh. BioMed Research International, 2017:6081470. Doi:10.1155/2017/6081470.

Ramesh, A., Blanchet, K., Ensink, J.H.J. and Roberts, B. (2015). Evidence on the effectiveness of water, sanitation, and hygiene (WASH) Interventions on Health Outcomes in Humanitarian Crises: A Systematic Review. PLoS ONE, 10 (9), pp.e0124688.

Ramirez, A., Ranis, G. and Stewart, F. (1998). Economic Growth and Human Development. Queen Elizabeth House Working Paper Series Number 18. University of Oxford, Oxford.

Rangel, J.M., Lopez, B., Mejia, M.A., Mendoza, C. and Luby, S. (2003). A novel technology to improve drinking water quality: a microbiological evaluation of in-home flocculation and chlorination in rural Guatemala. Journal of Water and Health 1, 15-22.

Rauniyar, G., Morales, A. and Melo, V. (2009). Impact of Rural Water Supply and Sanitation in Punjab, Pakistan. Impact evaluation study PAK 2009-26. Asian Development Bank (ADB), Islamabad.

Reese, H., Routray, P., Torondel, B., Sclar, G., Delea, M.G., Sinharoy, S.S., Zambrano, L., Caruso, B., Mishra, S.R., Chang, H.H. and Clasen, T. (2017). Design and rationale of a matched cohort study to assess the effectiveness of a combined household-level piped water and sanitation intervention in rural Odisha, India. BMJ Open 7:e012719. Doi:10.1136/bmjopen-2016-012719.

Reese, H., Routray, P., Torondel, B., Sinharoy, S.S., Mishra, S., Chang, H.H. and Clasen, T. (2019). Assessing longer-term effectiveness of a combined household-level piped water and sanitation intervention on child diarrhoea, acute respiratory infection, soil-transmitted helminth infection and nutritional status: a matched cohort study in rural Odisha, India. International Journal of Epidemiology, 48 (6), 1757-1767. Doi:10.1093/ije/dyz157.

Reeves, B., Wells, G.A. and Waddington, H. (2017). Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions – a taxonomy without labels. Journal of Clinical Epidemiology, 89, 30-42. Doi:10.1016/j.jclinepi.2017.02.016.

Reisch, J., Tyson, J. and Mize, S. (1989). Aid to the evaluation of therapeutic studies. Pediatrics, 84 (5), November.

Reller, M.E., Mendoza, C.E., Lopez, M.B., Alvarez, M., Hoekstra, R.M., Olson, C.A., Baier, K.G., Keswick, B.H. and Luby, S.P. (2003). A randomized controlled trial of household-based flocculant-disinfectant drinking water treatment for diarrhea prevention in rural Guatemala. American Journal of Tropical Medicine and Hygiene, 69 (4), 411-419.

Rhee, V., Mullany, L.C., Khatry, S.K., Katz, J., LeClerq, S.C., Darmstadt, G.L. and Tielsch, J.M. (2008). Impact of maternal and birth attendant hand-washing on neonatal mortality in southern Nepal. Archives of Pediatrics and Adolescent Medicine, 162 (7), 603-608.

Riley, R.D., Higgins, J.P.T. and Deeks, J.D. (2011). Interpretation of random effects meta-analysis. British Medical Journal, 342. https://doi.org/10.1136/bmj.d54.

Rogers, E. (2005). Diffusion of innovations. Fifth Edition. The Free Press, New York.

Root, G.P.M. (2001). Sanitation, community environments, and childhood diarrhoea in rural Zimbabwe. Journal of Health, Population and Nutrition, 19 (2), 73-82.

Rosenbaum, P.R. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70 (1), 41-55.

Rosenzweig, M. and Schultz, T.P. (1983). Estimating a household production function: heterogeneity, the demand for health inputs, and their effects on birth weight. Journal of Political Economy, 91 (5), 723-746.

Rosling, H., Rosling Rönnlund, A. and Rosling, O. (2018). Factfulness. Ten reasons we're wrong about the world – and why things are better than you think. Flatiron Books, New York.

Ross, I. (2019). Funders shouldn't misinterpret WASH-B and SHINE results as 'WASH doesn't work'. Blogpost, available at: https://washeconomics.com/2019/10/04/lets-not-misinterpret-wash-b-and-shine-results-as-wash-doesnt-work/ (accessed 18 February 2020).

Rothstein, H.R., Sutton, A.J. and Borenstein, M. (eds). (2005). Publication bias in meta-analysis: prevention, assessment and adjustments. John Wiley and Sons, Chichester. Doi:10.1002/0470870168.

Roushdy, R., Sieverding, M. and Radwan, H. (2011). The impact of water supply and sanitation on child health: evidence from Egypt. Population Council, Egypt Office. Available at: http://popcouncil.org/publications/wp.asp (accessed 1 September 2020).

Rubalcava, L., Teruel, G. and Thomas, D. (2009). Investments, time preferences and public transfers paid to women. Economic Development and Cultural Change, 57 (3), 507-538.

Rubin, D. (1974). Estimating causal effects of treatments in randomised and nonrandomised studies. Journal of Educational Psychology, 66 (5), 688-701.

Rubin, D. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science, 5 (4), 472-480.

Ryan, M.A.K., Christian, R. and Wohlrabe, J. (2001). Handwashing and respiratory illness among young adults in military training. American Journal of Preventative Medicine, 21, 79-83.

Ryder, R.W., Reeves, W.C., Singh, N., Hall, C.B., Kapikian, A.Z., Gomez, B., and Sack, R.B. (1985). The childhood health effects of an improved water supply system on a remote Panamanian island. The American Journal of Tropical Medicine and Hygiene, 34 (5), 921-924.

Sabet, S.M. and Brown, A. (2018). Is impact evaluation still on the rise? The new trends in 2010-2015. Journal of Development Effectiveness, 10 (3), 291-304.

Sacks, H., Chalmers, T.C. and Smith, H. (1982). Randomized versus historical controls for clinical trials. The American Journal of Medicine, 72, 233-240.

Samii, C., Lisiecki, M., Kulkarni, P., Paler, L. and Chavis, L. (2014). Effects of decentralized forest management (DFM) on deforestation and poverty in low and middle income countries: a systematic review. Campbell Systematic Reviews, 2014:10. Doi:10.4073/csr.2014.10.

Samii, C., Lisiecki, M., Kulkarni, P., Paler, L. and Chavis, L. (2014). Effects of payment for environmental services (pes) on deforestation and poverty in low and middle income countries: a systematic review. Campbell Systematic Reviews, 2014:11. Doi:10.4073/csr.2014.11.

Sańchez-Meca, J., Marin-Martinez, F. and Chacon-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. Psychological Methods, 8 (4), 448-467.

Sandieson, R. (2006). Pathfinding in the research forest: the pearl harvesting method for effective information retrieval. Education and training in developmental disabilities, 41 (4), 401-409.

Sansom, K., Franceys, R., Njiru, C. and Morales-Reyes, J. (2003). Contracting Out Water and Sanitation Services - Vol. 2. Case studies and analyses of Service and Management Contracts in developing countries, WEDC, Loughborough. Available at: https://wedc-knowledge.lboro.ac.uk/resources/books/Contracting_Out_Water_and_Sanitation_Services_-_Vol_2_-_Complete.pdf (accessed 18 February 2020).

Saran, A. and White, H. (2019). Evidence and gap maps: a comparison of different approaches. The Campbell Collaboration, Oslo. Doi.org/10.4073/cmdp.2018.2.

Sastry, N. (1996). Community characteristics, individual and household attributes, and child survival in Brazil. Demography, 33 (2), 211-229.

Saunders, R.J. and Warford, J.J. (1976). Village water supply: economics and policy in the developing world. A World Bank Research Publication. The John Hopkins University Press, Baltimore and London.

Savović, J., Jones, H., Altman, D., Harris, R., Jűni, P., Pildal, J., Als-Nielsen, B., Balk, E., Gluud, C., Gluud, L., Ioannidis, J., Schulz, K., Beynon, R., Welton, N., Wood, L., Moher, D., Deeks, J. and Sterne, J. (2012). Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technology Assessment, 16 (35), 1-82.

Scanlon, J.W. (1977). Evaluability assessment: Avoiding Type III or IV errors. In: G.R. Gilbert and P.J. Conklin (eds). Evaluation management: a source book of readings. U.S. Civil Service Commission, Charlottesville.

Schafer, J.L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychological Methods, 13 (4), 279-313.

Schmidt, F.L., Oh, I.-S. and Hayes, T.L. (2009). Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in result. British Journal of Mathematical and Statistical Psychology, 62, 97-128. Doi:10.1348/000711007X255327.

Schmidt, W.-P. (2014). Editorial: The elusive effect of water and sanitation on the global burden of disease. Tropical Medicine and International Health, 19 (5), 522-527.

Schmidt, W.-P. and Cairncross, S. (2009). Household water treatment in poor populations: Is there enough evidence for scaling up now? Environmental Science and Technology, 43 (4), 986-992.

Schmidt, W.-P. Cairncross, S., Barreto, M.L., Clasen, T. and Genser, B. (2009). Recent diarrhoeal illness and risk of lower respiratory infections in children under the age of 5 years. International Journal of Epidemiology, 38 (3), 766-772. Doi:10.1093/ije/dyp159.

Schmidt, W.-P., Arnold, B.F., Boisson, S., Genser, B., Luby, S.P., Barreto, M.L., Clasen, T. and Cairncross, S. (2011). Epidemiological methods in diarrhoea studies – an update. International Journal of Epidemiology, 40, 1678-1692. Doi:10.1093/ije/dyr152.

Schochet, P., Cook, T.D., Deke, J., Imbens, G., Lockwood, J.R., Porter, J. and Smith J. (2010). Standards for regression discontinuity designs. Mathematica Policy Research Report, Princeton, NJ.

Schumacher, E.F. (1973). Small is beautiful. A study of economics as if people mattered. Vintage Books, London.

Schünemann, H.J., Oxman, A.D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G.E., Williams, J.W., Kunz, R., Craig, J., Montori, V.M., Bossuyt, P., Guyatt, G.H. and GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. British Medical Journal, 336 (7653), 1106-10. Doi:10.1136/bmj.39500.677199.AE.

Scottish Intercollegiate Guidelines Network (SIGN) (2011). SIGN 50: A guideline developer's handbook. Revised Edition. Scottish Intercollegiate Guidelines Network, Edinburgh. Available at: http://www.sign.ac.uk/pdf/sign50nov2011.pdf [Accessed 1 December 2015].

Semenza, J.C., Roberts, L., Henderson, A., Bogan, J. and Rubin, C.H. (1998). Water distribution system and diarrheal disease transmission: a case study in Uzbekistan. The American Journal of Tropical Medicine and Hygiene, 59 (6), 941-946.

Sen, A. (1998). Mortality as an indicator of economic success and failure. Economic Journal, 108 (446), 1-25.

Shadish, W. and Cook, T. (2009). The renaissance of field experimentation in evaluating interventions. Annual Review of Psychology, 60, 607-29. Doi:10.1146/annurev.psych.60.110707.163544.

Shadish, W., Cook, T. and Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. BROOKS/COLE CENGAGE Learning.

Shadish, W.R., Clark, M.H. and Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. Journal of the American Statistical Association, 103, 1334-1344. Doi:10.1198/016214508000000733.

Shadish, W., Steiner, P.M. and Cook, T.D. (2012). A case study about why it can be difficult to test whether propensity score analysis works in field experiments. Journal of Methods and Measurement in the Social Sciences, 3, 1-12.

Shadish, W. (2013). Propensity score analysis: promise, reality and irrational exuberance. Journal of Experimental criminology, 9, 129-144.

Shaffer, P. (2013). Q-squared in impact assessment: a review. Q-Squared Working Paper No. 61, Winter 2012-2013.

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E. and Henry, D.A. (2016). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. British Medical Journal 2017;358:j4008 http://dx.doi.org/10.1136/bmj.j4008.

Shiffman, M.A., Schneider, R., Faigenblum, J.M. and Helms, R. (1978) Field studies on water sanitation and health education in relation to health status in Central America. Prog in Water Technology, 11 (1/2), 143-150.

Shor, E., Roelfs, D. and Vang, Z.M. (2017). The "Hispanic mortality paradox" revisited: meta-analysis and meta-regression of life-course differentials in Latin American and Caribbean immigrants' mortality. Social Science and Medicine, 186, 20-33.

Shuval, H.R., Tilden, R.L., Perry, B.H. and Grosse, R.N. (1981). Effect of investments in water supply and sanitation on health status: a threshold-saturation theory. Bulletin of World Health Organization, 59 (2), 243-248.

Siegel, J.S., Swanson, D.A. and Shryock, H.S. (2004). The methods and materials of demography. Second Edition. Elsevier/Academic Press, Amsterdam.

Sima, L.C., Desai, M. M., McCarty, K.M. and Elimelech, M. (2012). Relationship between use of water from community-scale water treatment refill kiosks and childhood diarrhea in Jakarta. American Journal of Tropical Medicine and Hygiene, 87 (6), 979-984. Doi:10.4269/ajtmh.2012.12-0224.

Skoufias, E., David, B. and de la Vega, S. (2001). Targeting the poor in Mexico: an evaluation of the selection of households into PROGRESA. World Development, 29 (10), 1769-1784.

Smith, J.C. and Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125, 303-353.

Smith, M. and Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32 (9), 752-760.

Snow, J. (1855). On the mode of communication of cholera. Second edition, much enlarged. John Churchill, London.

Sobel, J., Mahon, B., Mendoza, C.E., Passaro, D., Cano, F., Baier, K., Racioppi, F., Hutwagner, L. and Mintz, E. (1998). Reduction of fecal contamination of street-vended beverages in Guatemala by a simple system for water purification and storage, handwashing, and beverage storage. American Journal of Tropical Medicine and Hygiene, 59, 380-387.

Somers, M., Zhu, P., Jacob, R. and Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. MDRC Working Paper on Research Methodology.

Sommer, M., Ferron, S. Cavill, S. and House, S. (2014). Violence, gender and WASH: spurring action on a complex, underdocumented and sensitive topic. Environment and Urbanization, 27 (1), 105-116. Doi:10.1177/0956247814564528.

Sorenson, S. B., Morssink, C. and Campos, P. A. (2011). Safe access to safe water in low income countries: water fetching in current times. Social Science and Medicine, 72, 1522-1526.

Spears, D. (2013). Essays in the economics of sanitation and human capital in developing countries. PhD Thesis, Princeton, NJ.

Shrestha, S., Aihara, Y., Yoden, K., Yamagata, Z., Nishida, K. and Kondo, N. (2013). Access to improved water and its relationship with diarrhoea in Kathmandu Valley, Nepal: a cross-sectional study. BMJ Open, 2013;3:e002264. Doi:10.1136/bmjopen-2012-002264.

Stanton, B.F. and Clemens, J.D. (1987). An educational intervention for altering water sanitation behaviors to reduce childhood diarrhea in urban Bangladesh. American Journal of Epidemiology, 125 (2), 292-301.

Stanton, B.F., Clemens, J.D. and Khair, T. (1988). Educational intervention for altering water-sanitation behaviour to reduce childhood diarrhea in urban Bangladesh: impact on nutritional status. American Journal of Clinical Nutrition, 48 (5), 1166-1172.

Steiner, P.M. and Wong, V. (2016). Assessing correspondence between experimental and non-experimental results in within-study-comparisons. EdPolicy Works Working Paper Series No. 46, April 2016. University of Virginia, VA.

Stephenson, L.S., Latham, M.C. and Ottesen, E.A. (2000). Malnutrition and parasitic helminth infections. Parasitology, 2000;121 Suppl:S23-38.

Sterne, J.A.C., Hernán, M., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I, Carpenter, J.R., Chan, A.-W., Churchill, R., Deeks, J.J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F. and Higgins, J.P.T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. British Medical Journal, 2016;355:i4919. http://dx.doi.org/10.1136/bmj.i4919.

Sterne, J.A.C., Higgins, J.P.T. and Reeves, B.C. on behalf of the development group for ACROBAT- NRSI. (2014). A Cochrane risk of bias assessment tool: for non-randomized studies of interventions (ACROBAT- NRSI), Version 1.0.0, 24 September 2014. Available from: http://www.bristol.ac.uk/media-library/sites/social-community-medicine/images/centres/cresyda/ACROBAT-NRSI%20Version%201_0_0.pdf (accessed 26 June 2020).

Sterne, J.A.C., Jüni, P., Schulz, K.F., Altman, D.G., Bartlett, C. and Egger, M. (2002). Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in Medicine, 21, 1513-1524. Doi:10.1002/sim.1184.

Stewart, R., El-Harakeh, A. and Cherian, S.A. (2020). Evidence synthesis communities in low-income and middle-income countries and the COVID-19 response. Lancet, October 20, 2020, 1-2. Doi:10.1016/S0140-6736(20)32141-3.

Stocks, M.E., Ogden, S., Haddad, D., Addiss, D.G., McGuire, C. and Freeman, M.C. (2014). Effect of water, sanitation, and hygiene on the prevention of trachoma: a systematic review and meta-analysis. PLoS Medicine, 11 (2), e1001605.

Stoller, N.E., Gebre, T., Ayele, B., Zerihun, M., Assefa, Y., Habte, D., Zhou, Z., Porco, T.C., Keenan, J.D., House, J.I., Gaynor, B.D., Lietman, T.M. and Emerson, P.M. (2011). Efficacy of latrine promotion on emergence of infection with ocular Chlamydia trachomatis after mass antibiotic treatment: a cluster-randomized trial. International Health, 3 (2), 75-84.

Stone, R., de Hoop, T., Coombes, A. and Nakamura, P. (2019). What works to improve early grade literacy in Latin America and the Caribbean? A systematic review and meta-analysis. Campbell Systematic Reviews, 2019:16:e1067. Doi:10.1002/cl2.1067.

Stone, M.A. and Ndagijimana, H. (2017). Educational intervention to reduce disease related to sub-optimal basic hygiene in Rwanda: initial evaluation and feasibility study. Pilot and feasibility studies, 4 (4), 8 pages. https://doi.org/10.1186/s40814-017-0155-6.

Strachan, D.P. (2000). Family size, infection and atopy: the first decade of the 'hygiene hypothesis'. Thorax, 55 (Suppl 1), S2-S10.

Strina, A., Cairncross, S., Barreto, M.L., Larrea, C. and Prado, M.S. (2003). Childhood diarrhoea and observed hygiene behaviour in Salvador, Brazil. American Journal of Epidemiology, 157, 1032-38.

Strunz, E.C., Addiss, D.G., Stocks, M.E., Ogden, S., Utzinger, J. and Freeman, M.C. (2014). Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. PLoS Medicine, 11, pp.e1001620.

Sumpter, C. and Torondel, B. (2013). A systematic review of the health and social effects of menstrual hygiene management. PLoS ONE, 8 (4), April 26. Doi:10.1371/journal.pone.0062004.

Swanson, S.A., Tiemeier, H., Ikram, M.A. and Hernán, M.A. (2017). Nature as a trialist? Deconstructing the analogy between Mendelian randomization and randomized trials. Epidemiology, 28 (5), 653-659.

Tadesse, B., Worku, A., Kumie, A. and Yimer, S.A. (2017). Effect of water, sanitation and hygiene interventions on active trachoma in North and South Wollo zones of Amhara Region, Ethiopia: a quasi-experimental study. PLoS Neglected Tropical Diseases, 11, pp.e0006080.

Tanner, R.E.S. (1995). Excreting, excretors and social life: some preliminary observations on an unresearched activity. Journal of Preventative Medicine and Hygiene, 36, 85-94.

Tanner-Smith, E. and Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. Research Synthesis Methods, 5, 13-30. Doi:10.1002/jrsm.1091.

Terra de Souza, A.C., Cufino, E., Peterson, K., Gardner, J., Vasconcelos do Amaral, M.I. and Ascherio, A. (1999). Variations in infant mortality rates among municipalities in the state of Ceara, northeast Brazil: an ecological analysis. International Journal of Epidemiology, 28, 267-75.

Thomas, J., Brunton, J. and Graziosi, S. (2010). EPPI-reviewer 4: software for research synthesis. EPPI-Centre Software. Social Science Research Unit, UCL Institute of Education, London.

Thompson, J., Porras, I.T., Tumwine, J.K., Mujwahuzi, M.R,, Katui-Katua, M., Johnstone, N. and Wood, L. (2001). Drawers of Water II: 30 years of change in domestic water use and environmental health in east Africa. International Institute for Environment and Development, London.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics, 38 (3), 239-266. Doi:10.3102/1076998612441947.

Ton, G., Desiere, S., Vellema, W., Weituschat, S. and D'Haese, M. (2016). The effectiveness of contract farming for raising income of smallholder farmers in low- and middle-income countries: a systematic review. Campbell Systematic Reviews, 2017:13. Doi:10.4073/csr.2017.13.

Torún, B. (1982). Environmental and educational interventions against diarrhoea in Guatemala. In: Chen, L.C. and Scrimshaw, N.S. (eds). Diarrhea and malnutrition: interactions, mechanisms and interventions. Plenum Press, New York, NY.

Touré, O., Coulibaly, S., Arby, A., Maiga, F. and Cairncross, S. (2013). Piloting an intervention to improve microbiological food safety in Peri-Urban Mali, International Journal of Hygiene and Environmental Health, 216 (2), 138-145. https://doi.org/10.1016/j.ijheh.2012.02.003.

Trémolet, S. and Evans, B. (2010). Output-based aid for sustainable sanitation. OBA Working Paper Series Paper No. 10, September 2010. Global Partnership on Output-Based Aid, the World Bank, Washington, D.C.

Trent, M., Dreibelbis, R., Bir, A., Tripathi, S.N. Labhasetwar, P., Nagarnaik, P., Loo, A., Bain, R., Jeuland, M. and Brown, J. (2018). Access to household water quality information leads to safer water: a cluster randomized controlled trial in India. Environmental Science and Technology, 52 (9), 5319-5329. Doi:10.1021/acs.est.8b00035.

Tripney, J., Hombrados, J., Newman, M., Hovish, K., Brown, C., Steinka-Fry, K. and Wilkey, E. (2013). Technical and vocational education and training (TVET) interventions to improve the employability and employment of young people in low- and middle-income countries: a systematic review. Campbell Systematic Reviews, 2013:9. Doi:10.4073/csr.2013.9.

Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F. and Coiera, E. (2014). Systematic review automation technologies. Systematic Reviews, 3, 74, 1-15.

UN Water. (2018). Sustainable Development Goal 6 synthesis report on water and sanitation 2018. United Nations, New York.

UNICEF. (1990). Strategy for improved nutrition of children and women in developing countries. UNICEF, New York.

United Nations. (2015). The Millennium Development Goals report 2015. New York: United Nations.

United Nations. (undated). Report of the inter-agency and expert group on sustainable development goal indicators (E/CN.3/2016/2/Rev.1), Annex IV. Available at: https://sustainabledevelopment.un.org/content/documents/11803Official-List-of-Proposed-SDG-Indicators.pdf (accessed 28 January 2020).

Universidad Rafael Landívar. (1995). *Contra la morbilidad infantil: filtros artesanales y educación*. Estudios Sociales IV Epoca, 53, 1-66.

Vaessen, J., Rivas, A., Duvendack, M., Palmer-Jones, R., Leeuw, F.L., Van Gils, G., Lukach, R., Holvoet, N., Bastiaensen, J., Hombrados, J.G. and Waddington, H. (2014). The effects of microcredit on women's control over household spending in developing countries: a systematic review and meta-analysis. Campbell Systematic Reviews, 2014:8. Doi:10.4073/csr.2014.8.

Valentine, J. and Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device. Psychological Methods, 13 (2), 130-149.

Vander Hoorn, S., Ezzati, M., Rodgers, A., Lopez, A.D. and Murray, C.J.L. (2004). Estimating attributable burden of disease from exposure and hazard data. Chapter 25 in: Comparative Quantification of Health Risks. World Health Organization, Geneva, 2129–2140.

Venkataramanan, V., Crocker, J., Karon, A., and Bartram, J. (2018). Community-led total sanitation: a mixed-methods systematic review of evidence. Environmental Health Perspectives, 126 (2), 026001-1-17.

Verbeek, M. (2008). Pseudo-panels and repeated cross-sections. In: Matyas, L. and Sevestre, P. (eds). The econometrics of panel data. Springer-Verlag, Berlin Heidelberg.

Vickers, A., Goyal, N., Harland, R. and Rees, R. (1998). Do certain countries produce only positive results? A systematic review of controlled

trials. Controlled Clinical Trials, 19 (2), 159-166. Doi:10.1016/s0197-2456(97)00150-5.

Victora, C.G. and Barros, F.C. (2001). Infant mortality due to perinatal causes in Brazil: trends, regional patterns and possible interventions. Sao Paolo Medical Journal, 191 (1), 33-42.

Victora, C.G., Black, R.E., Boerma, J.T. and Bryce, J. (2011). Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. Lancet, 377, 85-95. Doi:10.1016/S0140-6736(10)60810-0.

Victora, C.G., Habicht, J.-P. and Bryce, J. (2004). Evidence-based public health: moving beyond randomized trials. American Journal of Public Health, 94 (3), 400-405.

Victora, C.G., Huttly, S.R., Fuchs, S.C. and Olinto, M.T. (1997) The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. International Journal of Epidemiology, 26 (1), 224-7. Doi:10.1093/ije/26.1.224.

Victora, C.G., Smith, P.G., Vaughan, J.P., Nobre, L.C., Lombardi, C., Teixeira, A.M.B., Fuchs, S.C., Moreira, L.B., Gigante, L.P. and Barros, F.C. (1988). Water supply, sanitation and housing in relation to the risk of infant mortality from diarrhoea. International Journal of Epidemiology, 17 (3), 651-654.

Vijayaraghavan, M., Kilroy, G., Jovellanos, J. and Regodon, C. (2018) Impact of cost-shared water supply services on household welfare in small towns. Ex-post impact evaluation of a project in Nepal. Impact Evaluation, May 2018. IES:NEP 2018-05. Asian Development Bank, Manila.

Villar, P.F. and Waddington, H. (2019). Within-study comparison and risk of bias in international development: systematic review and critical appraisal. Methods Research Paper, Campbell Systematic Reviews 2019;15:e1027. Doi:10.1002/cl2.1027.

Vist, G.E., Bryant, D., Somerville, L., Birminghem, T. and Oxman, A.D. (2009). Outcomes of patients who participate in randomised controlled trials compared to similar patients receiving similar interventions who do

not participate. Reprint. Cochrane Database of Systematic Reviews 2008, Issue 3. Art. No.: MR000009. Doi:10.1002/14651858.MR000009.pub4.

Viswanathan, S., Saith, R., Chakraborty, A., Purty, N., Malhotra, N., Singh, P., Mitra, P., Padmanabhan, V., Datta, S., Harris, J., Gidwani, S., Williams, R., Florence, E. and Sherin, D. (2019). Improving households' attitudes and behaviours to increase toilet use (HABIT) in Bihar, India. 3ie Grantee Final Report. International Initiative for Impact Evaluation, New Delhi.

Vivalt, E. (2018). Specification searching and significance inflation across time, methods and disciplines. Working Paper, Australian National University, October 25, 2018. Available at: http://evavivalt.com/wp-content/uploads/2015/09/Trajectory-of-Specification-Searching.pdf (accessed 11 May 2020).

Vivalt, E. (2020). How much can we generalize from impact evaluations? European Economic Review. Pre-publication version.

Wachter, K. (1988). Disturbed by meta-analysis? Science, 16 September, 241 (4872), 1407-1408, Doi:10.1126/science.3420397.

Waddington, H. (2014). Myths about microcredit and meta-analysis. Evidence Matters blogpost, 10 December 2014. International Initiative for Impact Evaluation, New Delhi.

Waddington, H. and Cairncross, S. (2020). Water, sanitation and hygiene (WASH) for reducing childhood mortality in low- and middle-income countries. Systematic Review Protocol. Campbell Systematic Reviews, in press.

Waddington, H. and Sabates-Wheeler, R. (2003). How does poverty affect migration choice? A review of literature. Working Paper T3, Development Research Centre on Migration, Globalisation and Poverty. University of Sussex, Brighton.

Waddington, H. and White, H. (2014). Farmer field schools: from agricultural extension to adult education. Systematic Review Summary 1. International Initiative for Impact Evaluation (3ie), London.

Waddington, H., Aloe, A.M., Becker, B.J., Djimeu, E.W., Hombrados, J.G., Tugwell, P., Wells, G. and Reeves, B. (2017). Quasi-experimental study

designs series-paper 6: risk of bias assessment. Journal of Clinical Epidemiology, 89, 43-52. Doi:10.1016/j.jclinepi.2017.02.015.

Waddington, H., Masset, E. and Jimenez, E. (2018). What have we learned after ten years of systematic reviews in international development? Journal of Development Effectiveness, 10 (1), 1-16.

Waddington, H., Snilstveit, B., Hombrados, J., Vojtkova, M., Phillips, D., Davies, P. and White, H. (2014). Farmer field schools for improving farming practices and farmer outcomes: a systematic review. Campbell Systematic Reviews, 2014:6. Doi:10.4073/csr.2014.6.

Waddington, H., Snilstveit, B., White, H. and Fewtrell, L. (2009). Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries. Synthetic Review 01. 3ie, New Delhi.

Waddington, H., Sonnenfeld, A., Finetti, J., Gaarder, M., John, D. and Stevenson, J. (2019). Citizen engagement in public services in low- and middle-income countries: a mixed-methods systematic review of participation, inclusion, transparency and accountability (PITA) initiatives. Campbell Systematic Reviews, 2019;15:e1025. Doi:10.1002/cl2.1025.

Waddington, H., White, H., Snilstveit, B., Hombrados, J.G., Vojtkova, M., Davies, P., Bhavsar, A., Eyers, J., Perez Koehlmoos, T., Petticrew, M., Valentine, J.C. and Tugwell, P. (2012). How to do a good systematic review of effects in international development: a toolkit. Journal of Development Effectiveness, 4 (3), 359-387.

Wagner, E.G., and Lanoix, J.N. (1958). Excreta disposal for rural areas and small communities. World Health Organization Monograph Series No. 39. WHO, Geneva.

Wagner, E.G., and Lanoix, J.N. (1959). Water supply for rural areas and small communities. World Health Organization Monograph Series No. 42. WHO, Geneva.

WaterAid, undated. Nowhere to go: how a lack of safe toilets threatens to increase violence against women in slums. Available at: www.wateraid.org/se/~/media/Files/Sweden/nowhere-to-go.pdf

Waterkeyn, J. and Cairncross, S. (2005). Creating demand for sanitation and hygiene through community health clubs: a cost-effective intervention in two districts in Zimbabwe. Social Science and Medicine, 61, 1958-1970.

Watson, J.A., Ensink, J.H.J., Ramos, M., Benelli, P., Holdsworth, E., Dreibelbis, R. and Cumming, O. (2017). Does targeting children with hygiene promotion messages work? The effect of handwashing promotion targeted at children, on diarrhoea, soil-transmitted helminth infections and behaviour change, in low- and middle-income countries. Tropical Medicine and International Health, 22, 526-538.

Welch, V.A., Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z.A., Cumberbatch, C., Fletcher, R., McGowan, J., Krishnaratne, S., Kristjansson, E., Sohani, S., Suresh, S., Tugwell, P., White, H. and Wells, G. (2017). Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis. Lancet Global Health, 5 (1), e40-e50.

Wells, G. (Undated). Newcastle-Ottawa quality assessment scale. Mimeo.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F. and Lux, L. (2002). Systems to rate of strength of scientific evidence. Evidence Report/Technology Assessment Number 47. AHRQ Publication No. 02-E016.

White, G.R., Bradley, D.J. and White, A.U. (1972). Drawers of water: domestic water use in East Africa. University of Chicago Press, IL.

White, H. (2004). Reducing infant and child death. Chapter 10 in: Black, R. and White, H. (eds). Targeting development: critical perspectives on the Millennium Development Goals. Routledge, London.

White, H. (2009). Theory-based impact evaluation. Journal of Development Effectiveness, 1 (3), 271-284.

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. Journal of Development Effectiveness, 5 (1), 30-49. Doi:10.1080/19439342.2013.764652.

White, H. (2014). Current challenges in impact evaluation. European Journal of Development Research, 26 (1), 18-30.

White, H. (2018). Theory-based systematic reviews. Journal of Development Effectiveness, 10 (1), 17-38. https://doi.org/10.1080/19439342.2018.1439078.

White, H. and Gunnarsson, V. (2008). What works in water supply and sanitation? Lessons from impact evaluations. A summary of findings. June 30, 2008. Independent Evaluation Group, the World Bank, Washington, D.C.

White, H. and Waddington, H. (2012). Why do we care about evidence synthesis? An introduction to the special issue on systematic reviews. Journal of Development Effectiveness, 4 (3), 1-12.

White, H., Blöndal, N., Masset, E. and Waddington, H. (2005). Maintaining momentum to 2015? An impact evaluation of interventions to improve maternal and child health and nutrition in Bangladesh. Operations Evaluation Department, the World Bank, Washington, D.C.

White, H., Menon, R. and Waddington, H. (2018). Community-driven development: does it build social cohesion or infrastructure? A mixed-method evidence synthesis. 3ie Working Paper 30. 3ie, New Delhi.

Whittington, D. (2002). Improving the performance of contingent valuation studies in developing countries. Environmental and Resource Economics, 22, 323-367.

Whitty, C.J.M. (2015). What makes an academic paper useful for health policy? BMC Medicine, 13, 301. https://doi.org/10.1186/s12916-015-0544-8.

WHO / UNICEF. (2013). Progress on sanitation and drinking-water: 2013 update. UNICEF and WHO, Geneva.

WHO / UNICEF. (2017). Progress on drinking water, sanitation, and hygiene: 2017 update and SDG baselines. Geneva, Switzerland: UNICEF and WHO.

WHO / UNICEF. (2019). Progress on drinking water, sanitation, and hygiene 2000-2017: special focus on inequalities. Geneva, Switzerland: WHO / UNICEF.

WHO. (1983). Minimum evaluation procedure. World Health Organization, Geneva.

WHO. (2017). UN-Water global analysis and assessment of sanitation and drinking-water GLAAS 2017 Report: Financing universal water, sanitation and hygiene under the sustainable development goals. WHO, Geneva.

WHO. (2018). Global health estimates 2016: global health estimates 2016: disease burden by cause, age, sex, by country and by region, 2000-2016. World Health Organization, Geneva. Available at: https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html (accessed 28 August 2020).

WHO/UNICEF. (2000). Global water supply and sanitation assessment 2000 report. Water Supply and Sanitation Collaborative Council. World Health Organization/United Nations Children's Fund, New York.

Winter, S.C. and Barchi, F. (2016). Access to sanitation and violence against women: evidence from Demographic Health Survey (DHS) data in Kenya. International Journal of Environmental Health Research, 26, 291-305. 10.1080/09603123.2015.1111309.

Wolf, J., Hunter, P.R., Freeman, M.C., Cumming, O., Clasen, T., Bartram, J., Higgins, J.P.T., Johnston, R., Medlicott, K., Boisson, S. and Prüss-Ustün, A. (2018). Impact of drinking water, sanitation and handwashing with soap on childhood diarrhoeal disease: updated meta-analysis and meta-regression. Tropical Medicine and International Health, 23 (5), 508-525.

Wolf, J., Johnston, R., Hunter, P.R., Gordon, B., Medlicott, K. and Prüss-Ustün, A. (2019). A Faecal Contamination Index for interpreting heterogeneous diarrhoea impacts of water, sanitation and hygiene interventions and overall, regional and country estimates of community sanitation coverage with a focus on low- and middle-income countries. International Journal of Hygiene and Environmental Health, 222, 270-282.

Wolf, J., Prüss-Ustün, A., Cumming, O., Bartram, J., Bonjour, S., Cairncross, S., Clasen, T., Colford, J.M., Curtis, V., De France, J., Fewtrell, L., Freeman, M.C., Gordon, B., Hunter, P.R., Jeandron, A., Johnston, R.B., Mäusezahl, D., Mathers, C., Neira, M. and Higgins, J.P. (2014). Assessing the impact of drinking water and sanitation on diarrhoeal

disease in low- and middle-income settings: systematic review and meta-regression. Tropical Medicine and International Health, 19 (8), 928-42. Doi:10.1111/tmi.12331. Epub 2014 May 8.

Wong, V. and Steiner, P. (2016). Designs of empirical evaluations of non-experimental methods in field settings. EdPolicy Works Working Paper Series No. 44, April 2016. University of Virginia, VA.

Wong, V., Valentine, J. and Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. Journal of Research on Educational Effectiveness, 10 (1), 207-236.

Wood, L., Egger, M., Gluud, L.L., Schulz, K.F., Jüni, P., Altman, D.G., Gluud, C., Martin, R.M., Wood, A.J.G. and Sterne, J.A.C. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. British Medical Journal, 2008; 336:601. Doi:10.1136/bmj.39465.451748.AD.

Wooldridge, J.M. (2009). Introductory econometrics: a modern approach. Fourth Edition. South-Western CENGAGE Learning, Canada.

World Bank (2017). Pipe(d) dreams: water supply, sanitation, and hygiene. Progress and remaining challenges in Ecuador. WASH Poverty Diagnostic. The World Bank, Washington, D.C.

World Bank Operations Evaluation Department (OED). (1998). Paraguay impact evaluation report: community-based rural water systems and the development of village committees. Report No. 17923. The World Bank, Washington, D.C.

World Bank. (2008). Economic impacts of sanitation in Southeast Asia: a four-country study conducted in Cambodia, Indonesia, the Philippines and Vietnam under the Economics of Sanitation Initiative (ESI). Water and Sanitation Program, the World Bank, Washington, D.C.

Wright, G., Coudert, F.X., Bentley, M., Steel, G. and Deville, S. (2014). This study is intentionally left blank. A systematic literature review of blank pages in academic publishing. Available at: https://figshare.com/articles/journal_contribution/This_Study_is_Intent

ionally_Left_Blank_A_systematic_literature_review_of_blank_pages_in_academic_publishing/1230110 (accessed 3 August 2021).

Wright, J., Gundry, S. and Conroy, R. (2004). Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. International Health, 9 (1), 106. ISSN 13602276. ISBN 13602276.

Yates, T., Allen, J., Joseph, M.L. and Lantagne, D. (2017). Short-term WASH interventions in emergency response: a systematic review. Systematic Review 33. 3ie, London.

Yeager, B.A.C., Lanata, C.F., Lazo, F., Verastegui, H. and Black, R.E. (1991). Transmission factors and socio-economic status as determinants of diarrhoeal incidence in Lima, Peru. Journal of Diarrhoeal Disease Research, 9, 186-93.

Zafar, S.N., Libuit, L., Hashmi, Z.G., Hughes, K., Greene, W.R., Cornwell III, E.E., Haider, A.H., Fullum, T.M. and Tran, D.D. (2015). The sleepy surgeon: does night time surgery for trauma affect mortality outcomes? The American Journal of Surgery, 209 (4), 633-639. Doi:10.1016/j.amjsurg.2014.12.015.

Zaheer, M., Prasad, B.G., Govil, K.K. and Bhadury, T. (1962). A note on urban water supply in Uttar Pradesh. Journal of the Indian Medical Association, 38, 177-82.

Zhang, J. (2011). The impact of water quality on health in rural china. Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Ziegelbauer, K., Speich, B., Mäusezahl, D., Bos, R., Keiser, J. and Utzinger, J. (2012). Effect of sanitation on soil-transmitted helminth infection: systematic review and meta-analysis. PLoS Med, 9 (1), e1001162. Doi:10.1371/journal.pmed.1001162.

Ziegelhöfer, Z. (2012). Down with diarrhea - using fuzzy regression discontinuity design to link communal water supply with health, IHEID

Working Papers 05-2012, Economics Section, The Graduate Institute of International Studies.

Zwane, A.P., Zinman, J., van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D., Hornbeck, R., Gine, X., Duflo, E., Devoto, F., Crepon, B. and Banerjee, A. (2011). Being surveyed can change later behavior and related parameter estimates, Proceedings of the National Academy of Sciences of the United States of America, 108 (5), 1821-1826.

Appendix A Critical appraisal tool for randomised and non-randomised studies of effects

Table A1 Study design and external validity

Question	Description
Unique study ID	
Study first author, year	Open answer
Outcome	Open answer: this checklist should be completed for each outcome separately, in the case of multiple reported outcomes.
Intervention: describe the intervention and mechanism	• Clarify whether the independent variable in the study measures provision of an **intervention** or an **exposure** to a technology. • Clarify whether the intervention is **baseline/point** (e.g., administration of deworming tablet) or **continuous** (e.g., WASH technology intervention requiring sustained behaviour change). • Clarify the possible **intervention mechanism(s)** through which the treatment effect operates.
Allocation of treatment: describe how the treatment was assigned	• Indicate whether the intervention is allocated by researchers (e.g., through randomisation, discontinuity assignment, statistical matching), policy-makers/practitioners (e.g., through a lottery, individual/household means-testing, community/geographic targeting), or participants (through self-selection), or a combination? • Indicate what information is known about the **intervention allocation** mechanism at group and individual levels in the study (e.g., if allocated by decision-makers, what allocation rules are used). • Clarify **unit of randomisation** (unit of random assignment to treatment and control by researchers, if relevant), **unit of treatment** (unit of intervention) and **unit of analysis** (data collection unit).
External validity: describe the intervention design and implementation, sampling,	• Indicate who **designed and implemented the intervention** (whether by researchers, policymakers and/or practitioners). • Clarify the **intervention scale**: whether the study is a trial, pilot study or small-scale project (e.g., implemented in a few villages by researchers), or an evaluation of a scaled-up programme (e.g., implemented at province or national level by government, private sector or NGO).

survey design, and use of explicit theory	• Indicate the **intervention period** length and information about within-intervention period follow-up (how many visits, how often, by whom, for what purpose). • Clarify the **sampling frame** for the data collection, and sampling approach at cluster and individual levels (whether random or purposive). • Indicate the **study length** (follow-up period) and number of follow-ups (outcomes data collection points, including baseline if relevant). • Specify whether there is an explicit **programme theory** presented in the paper (e.g., a logic model showing the causal pathway) • Specific whether data are collected on outputs, intermediate outcomes and endpoint outcomes (**causal pathway analysis**).
Implementation: describe how implementation and adherence were measured	• Specify any information about **implementation fidelity**; methods of assessing implementation fidelity; results of the assessment. • Specify whether any information is given about **programme take-up** (among participants); methods of assessing take-up; results of the assessment. • Specify any information about **adherence** (by participants); methods of assessing adherence; results of the assessment.
Study design RCT: what type of study design is used?	1= Randomised controlled trial (RCT) (random assignment of individuals to intervention) 2= Cluster-RCT (random assignment of groups to intervention) 3= Quasi-RCT (e.g., prospective assignment to intervention by alternation of individuals or group ordered by alphabet)
Study design NRS: what type of study design is used?	1= Natural experiment: randomised or quasi-randomised (e.g., 'as-if' randomisation by implementation error) 2= Regression discontinuity design (RDD) or geographical discontinuity design (GDD) 3= Interrupted time series (ITS), or controlled-ITS with contemporaneous comparison group 4= Instrumental variables (IV) study: e.g., randomised encouragement to universal programme 5= Panel study (individual repeated measurement): non-randomised assignment with pre-test and post-test outcomes data collection in treatment and comparison 6= Pseudo panel study: repeated measurement of outcomes at pre-test and post-test for groups but different individuals 7= Post-test panel: repeated outcomes data collection in treatment and comparison, but no outcomes data collection at pre-test (e.g., cohort study) 8= Case-control study: outcomes data collection in treatment and comparison group at post-test, where cases (those experiencing the outcome) are matched to 'controls' (those who do not experience the outcome)

	9= Cross-section study: data collection in treatment and comparison group at a single point in time post-test, where the relationship between outcomes and characteristics of individuals or groups is assessed 10= Pre-test and post-test data collection in treatment group only (before versus after study) 11= Post-test data collection in treatment group only (single case design)	
Treatment estimand and methods of analysis:	1= Intention-to-treat (ITT), reduced form unadjusted estimation or comparison of group means 2= ITT, covariate-adjusted estimation 3= ITT, with fixed effects 4= ITT, with double differences (DD) estimation 5= Complier average causal effect (CACE) using IV estimation 6= Local average treatment effect (LATE) using a sub-sample of observations around a treatment threshold (RDD or GDD) 7= Average treatment effect on the treated (ATET), as estimated typically by statistical matching (e.g., propensity score matching, PSM), also called treatment-on-the-treated (TOT) or the per-protocol effect in an RCT. 8= ATET with statistical matching on baseline outcome or DD estimation (pre-test and post-test outcome data) 9= Other (indicate)	
Design and method description	Open answer	Briefly describe the study design and analysis method undertaken by the authors.
Blinding of participants Were participants blinded to treatment status?	Y, N, U	If there is no information, code N. If there is information but it is ambiguous, code U.
Blinding of observers Were outcome assessors blinded to treatment status?	Y, N, U	If there is no information, code N. If there is information but it is ambiguous, code U.
Blinded analysts Were data analysts blinded to treatment status?	Y, N, U	If there is no information, code N. If there is information but it is ambiguous, code U.
Method used to blind	Open answer	Describe method(s) used to blind including method for placebo control.

Table A2 Risk-of-bias assessment signalling questions and decision rules

Bias domain	Question	Coding	Scoring criteria	Decision rules
1	**1a. Confounding**: Was the allocation or identification mechanism able to address confounding?			
	• **RCT**	Y, PY, PN, N, U	**a) Sequence generation:** **-** The authors describe a random component in sequence generation/ randomisation method (e.g., lottery, coin toss, random number table).* **-** If a special randomisation procedure is used to ensure balance, it is well described (stratification, pairwise matching, unique random draw, multiple random draws etc.) and adjustment is considered in the analysis (e.g., stratum fixed effects, pairwise matching variables). **b) Subversion:** - if the unit of allocation was by beneficiary or group, there was some form of centralised allocation mechanism such as an on-site computer system to ensure adequate allocation concealment. **-** If a public lottery was used for the sequence generation, details were given on the exact settings and participants attending the lottery. **c) Balance:** **-** The unit of allocation is based on a sufficiently large sample size to equate groups on average.	- Score “Low risk” if all criterion are satisfied. - Score "Some concerns" if there is no balance table reported (or key variables are omitted from the table) -- Score "High risk" if there is any failure in the allocation mechanism which could affect the randomisation process, or there is no balance table reported (c) and there is evidence suggesting a problem in the randomisation, such as covariate means are very different or sample size is too small for the procedure used (using stratification when there are less than two units for each intervention and control group in each strata can lead to imbalance), or if the paper does not provide details on the randomisation process or uses quasi-randomisation (e.g., alternate households allocated) which it is not clear has generated allocations equivalent to randomisation. * In order to assess the validity of the quasi-randomisation process, the most important

Bias domain	Question	Coding	Scoring criteria	Decision rules
			- A balance table is reported for all subgroups receiving differential treatment, comparing means and standard deviations of variables, including cluster-level variables.	aspect is whether the assignment process might generate a correlation between participation status and other factors (for example, gender, socio-economic status, pre-existing health condition) determining outcomes; consider whether assignment is done at cluster level (centralised) and covariate balance is reported.
	• **Discontinuity design**	Y, PY, PN, N, U	**a) Allocation:** information about the programme targeting criteria are known, presented in the paper, and used to justify the statistical approach. Demonstration of the relationship between the assignment variable (a continuous variable or a discrete scaled variable with sufficient points either side of the cut-off) and outcome is done using a graph of the assignment-outcome relationship. Appropriate functional form may include local linear regression at assignment threshold or ordered polynomial. The treatment effect may be measured as a change in intercept and/or change in slope. **b) Subversion:** Classification of intervention status is not affected by systematic manipulation of the assignment variable by participants or decision-makers, as indicated by: - the assignment decision rule is concealed from participants and practitioners, or - the assignment variable is non-manipulable by participants, practitioners or other decision-makers, or	-Score “Low risk” if all criteria are satisfied. -Score "Some concerns" if participants or practitioners are unblinded or confirmation or falsification tests suggest potential problems. -Score “High risk” if there are important differences between individuals on both sides of the cut-off, and confirmation or falsification tests suggest potential problems, or if confirmation or falsification tests are not reported.

Bias domain	Question	Coding	Scoring criteria	Decision rules
			- the assignment variable is measured with random error. To verify this, the study should report a histogram of the assignment variable to demonstrate that bunching does not occur around the threshold, and McCrary's (2006) test should be reported. **c) Confirmation/falsification:** The relationship between assignment variable and outcomes are unconfounded at the threshold. Support for this can be obtained by confirmation test of no discontinuity at the cut-off in terms of baseline characteristics around the threshold, and falsification tests such as: - Addition of a phase in which intervention is not present, or 'placebo time period', e.g., by estimating the pre-test relationship between assignment variable and outcomes, as a falsification exercise. Responsiveness of the outcome variable to temporal changes in intervention can also help verify the functional form and to adjust for non-linearities in the relationship. - Addition of a non-equivalent outcome, or 'placebo outcome'; that is, assessing the effect on a second outcome variable that the intervention should not influence, as a falsification exercise. - Use of 'placebo discontinuity' tests showing no other discontinuities in the assignment variable within the bandwidth of interest, as a falsification exercise.	

Bias domain	Question	Coding	Scoring criteria	Decision rules
	• **NRS using statistical matching**	Y, PY, PN, N, U	**a)** Information about the programme targeting criteria are known, presented in the paper, and used to justify the statistical approach. **b)** Matching is done on pre-test (or time-invariant) characteristics, including the outcome measured at pre-test; matches are geographically local; the variables used to match are relevant (for example, demographic and socio-economic factors) to explain both participation and the outcome (so that there can be no evident differences across groups in variables that might explain outcomes); and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme.* **c)** With the exception of Kernel matching, the means of the individual covariates are demonstrated to be equated for treatment and comparison groups after matching.	-Score "Low risk", if all criteria are addressed. -Score "Some concerns " if the selection into the programme was done according to clear targeting rules, which are used as matching variables, but there are imbalances remaining after matching. -Score "High risk" if programme assignment was self-selected by participants and no baseline data are available to match the participants or groups, or matching was done based on variables that are likely to be affected by the programme, or relevant variables are not included in the matching equation including cluster-level variables. * Accounting for and matching on all relevant characteristics is usually only feasible when the programme allocation rule is known and there are no errors of targeting. There are different ways in which covariates can be considered. Observable differences across groups can be incorporated in the framework of a regression analysis (e.g., propensity-weighted least squares) or can be assessed by testing equality of means between groups. Differences in unobservable characteristics can be account for using double differences (DD), fixed effects (FE) or

Bias domain	Question	Coding	Scoring criteria	Decision rules
				random effects (RE) where unobservables are time-invariant.
	• **NRS using double differences (DD), fixed effects (FE) or random effects (RE) analysis of panel data***	Y, PY, PN, N, U	**a)** Outcomes are measured at pre-test (before intervention) and post-test (after intervention) using the same approach. **b)** Examination of secular trends in outcomes shows parallel trends across treatment and comparison groups during periods prior to intervention. **c)** The method is combined by well-conducted statistical matching done according to clear programme allocation rules (see above), and baseline imbalances, including in the outcome are shown to be small. **d)** A comprehensive set of individual time-varying characteristics is controlled, including any cluster-level covariates that may affect the impact of the programme (e.g., rainfall).**	-Score "Low risk" all criteria are addressed. -Score "Some concerns" if selection into the programme was done according to clear rules, and equal trends demonstrated, but baseline imbalances between groups remained. -Score "High risk " if equal trends are not reported, and programme allocation was due to participant self-selection, programme allocation was self-selected by participants and some relevant time-varying characteristics are not controlled, or insufficient details are provided, for example on testing the equal trends assumption or about cluster-level variables. * DD, FE and RE regression models are sometimes complemented with matching strategies. This combination approach is superior since it only uses in the estimation the common support region of the sample size, reducing the likelihood of existence of time-varying unobservable differences across groups affecting outcome of interest and removing biases arising from time-invariant unobservable characteristics. ** Knowing allocation rules for the programme – or even whether the non-

Bias domain	Question	Coding	Scoring criteria	Decision rules
				participants were individuals that refused to participate in the programme, as opposed to individuals that were not given the opportunity to participate in the programme – can help in the assessment of whether the covariates accounted for in the regression capture all the relevant characteristics that explain differences between treatment and comparison.
	• **Instrumental variables (IV) estimation**	Y, PY, PN, N, U	**a)** An appropriate instrumental variable is used which is exogenously generated: for example, due to a 'natural' experiment or random allocation. If the instrument is the random assignment of the treatment, or fuzzy discontinuity, the reviewer should also assess the randomisation procedure or discontinuity assignment, as above. **b)** The joint test for the instruments is significant at the level of F≥10, or if an F test is not reported, the authors report and assess whether the R-squared (goodness of fit) of the participation equation is sufficient for appropriate identification; and the identifying instruments are individually significant (p≤0.01). **c)** The study assesses qualitatively why the instrument only affects the outcome via participation (the exclusion restriction); where at least two instruments are used, the authors report on an over-identifying test (p≤0.05 is required to reject the null hypothesis); and none of the covariate controls can be	-Score "Low risk", if all criteria are addressed. -Score "Some concerns" if tests required for criterion b) are not satisfied, but the rest of the criterion are addressed and the exogeneity of the instrument is clear. -Score "High risk" if exogeneity of the instrument is not convincing and appropriate tests are not reported, or if insufficient details are provided on cluster controls.

Bias domain	Question	Coding	Scoring criteria	Decision rules
			affected by participation. **d)** Authors control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth).	
	1b. Confounding - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub-questions, cite relevant pages).	
2	**2a. Selection bias:** was any differential selection into the study adequately resolved?	Y, PY, PN, N, U	**a) Follow-up data:** If the study design is prospective, follow-ups are recorded for all eligible participant units from recruitment onwards (i.e., prior to treatment). This is best shown using a participant flow diagram or reporting sufficient information to construct one. **b) Participant identification:** where the unit of allocation in a prospective study was at group level (geographical/ social/ cluster unit), allocation was performed on all units at the start of the study, or participants and recruiters are blinded to allocation status, or awareness is unlikely to affect recruitment differentially (e.g., participants chosen randomly using a sampling frame based on census and response rate is high). **c) Balance:** a table is reported for all subgroups receiving differential treatment within control or treatment groups, comparing means and standard deviations of variables; any unbalanced covariates at individual level are controlled in adjusted analysis, including cluster-level variables.	-Score "Low risk" if all relevant criteria are satisfied. -Score "Some concerns" if the study used prospective design with adequate concealment, but no (or an incomplete) study flow diagram is reported, or in retrospective design where statistical methods are used to correct for selection bias. -Score "High risk" if there are threats to adequate concealment (e.g., individual participants were chosen after cluster assignment was conducted or known, and there are differences between characteristics of the two groups beyond those expected by chance alone), or there is evidence of differential recruitment into study arms and differences in characteristics of groups not compatible with chance, or if no information is presented about participant characteristics or, in a prospective study, no study flow diagram (or data to construct it) presented.

Bias domain	Question	Coding	Scoring criteria	Decision rules
			d) Selection bias analysis: where evidence suggests there is selection bias into the study due to censoring of data (e.g., immortal time bias), this is accounted for using appropriate statistical methods (e.g., propensity weighted regression, Heckman selection model, proportional hazards model).	
	2b. Selection bias - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub-questions, cite relevant pages).	
3	**3a. Attrition bias:** was any differential selection out of the study adequately resolved?	Y, PY, PN, N, U	**a) Attrition at cluster-level:** is sufficiently low and similar reasons for attrition in treatment and control. Sufficiently low attrition is defined as: - total attrition (losses to follow-up) between pre-test and post-test in the study less than 10 percent of clusters (low risk) or 20 percent (some concerns). - differential cluster attrition across study arms is less than 10 percentage points, and reasons for attrition are given and similar across groups. **b) Attrition at individual-level:** is sufficiently low and similar reasons for attrition in treatment and control. Sufficiently low attrition is defined as: - total attrition (losses to follow-up) between pre-test and post-test in the study less than 10 percent of observations (low risk) or 20 percent (some concerns). - differential attrition across study arms is less than 10 percentage points, and reasons for attrition are given and similar across groups. **c) Robustness to attrition:** the study assesses losses to follow-up to be random draws from the	-Score "Low risk" if overall attrition is less than 10 percent and differential attrition less than 10 percentage points at cluster (a) and individual (b) levels, and the study demonstrates robustness to attrition. -Score "Some concerns" if overall attrition is between 10% and 20% and differential attrition less than 10 percentage points. -Score "High risk" if overall attrition exceeds 20% or differential attrition exceeds 10 percentage points, or there is some indication that the survey respondents were purposively sampled in a way that might have led the sampling to be different between treatment and control groups, or there is insufficient information on sampling methods, or no information on attrition is given.

Bias domain	Question	Coding	Scoring criteria	Decision rules
			sample (for example, by examining correlation with key characteristics across groups, or an F-test of attrition on baseline characteristics and interacted with treatment status), and study participants are randomly sampled.	
	3b. Attrition bias - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
4	**4a. Motivation bias:** was the process of observation free from motivation bias?	Y, PY, PN, N, U	**Are criteria adequately addressed?** **a)** For data collected in the context of a particular intervention trial (randomised or non-randomised assignment), the authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded to participants and outcome assessors, or methods are used that would minimise risk of Hawthorne effects, John Henry effects or survey effects such as infrequent observation or outcome questionnaires not referring to the intervention. Authors may also adapt the study design to estimate possible survey and Hawthorne effects (e.g., a 'pure control' with no monitoring except baseline endline). **b)** Informed consent is not associated with a particular intervention, as in the case of a regular household survey or a cluster-RCT, data are collected from administrative records, or in the context of a retrospective (*ex post*) evaluation.	-Score "Low risk" if either criterion is satisfied. -Score "Some concerns" if there was imbalance in the frequency of monitoring in intervention groups, which could have influenced behaviour in treatment and control differentially. - Score "High risk" if authors do not use an appropriate method to prevent possible motivation biases through blinding or other controls (e.g., infrequent measurement, methods to ensure consistent monitoring across groups, measurement using a 'pure control').

Bias domain	Question	Coding	Scoring criteria	Decision rules
	4b. Motivation bias - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
5	**5a. Performance bias:** was the study adequately protected against spillovers, no-shows and crossovers?	Y, PY, PN, N, U	**a)** There were no implementation issues that might have led the control participants to receive the treatment, or authors use intention-to-treat (ITT) estimation. **b)** The intervention is unlikely to spill over to comparisons (e.g., participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are not likely), or the potential effects of spillovers were measured (e.g., variation in the % of units within a cluster receiving the treatment). **c)** There is no risk of substitution (differential contamination) by external programs (also called treatment confounding): participants are isolated from other interventions which might be received differentially between treatment and controls which could explain changes in outcomes. **d)** Errors in implementation fidelity by the intervening body were not systematic, or unlikely to affect the outcome. **e)** For continuous interventions, measurement is taken of adherence to treatment among participants.	-Score “Low risk” if all criteria are satisfied. -Score "Some concerns" if there is no obvious problem but there is no information reported on potential risks related to spillovers or contamination in the control group, or if there were issues with spillovers but they were controlled for or measured, or if any of the criteria are not satisfied but the scale of the issue is minimal. -Score “High risk” if any of the criterion are not satisfied and happened at a large scale in the study, or if spillovers, no-shows, crossovers, implementation fidelity, or adherence to continuous interventions, are not reported clearly.
	5b. Deviation from interventions - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

Bias domain	Question	Coding	Scoring criteria	Decision rules
6	**6a. Measurement error:** is the study free from biases in measurement of intervention and outcomes?	Y, PY, PN, N, U	**a)** The study is a prospective design or in a retrospective design, participation in the intervention is observed, or the intervention clearly and consistently defined and misreporting by participants or enumerators is unlikely. **b)** Outcomes are clearly and consistently defined for all participants and outcome assessors in the study. **c)** Outcomes are measured through observation (rather than self-report), and outcome assessors are blinded to intervention or it is shown they are unbiased (e.g., spot-checks to validate). **d)** For self-reported outcomes: respondents in the intervention group are not more likely to report accurately than controls due to recall bias. **e)** Respondents do not have incentives to over/under report something related to their performance or actions, or researchers put in place mechanisms to reduce the risk of reporting bias (irregular or infrequent data collection rounds, outcome assessors not involved in the implementation of the intervention, it is clear that answers to the survey will not affect what they receive in the future), or authors have measured bias through falsification tests (e.g., 'placebo outcomes' in cases where there was a risk of reporting bias). **f)** Timing of the data collection did not differ between intervention and comparison group, the baseline data are not likely to be differentially affected by the time of intervention (e.g., due to seasonality).	-Score "Low risk" if all criteria are satisfied. -Score "Some concerns" if there is a small risk related to any criteria and potential biases are measured, e.g., with placebo outcomes, and found to be null. -Score "high risk" if there are risks related to any criteria and authors were not able to control for the bias, or no information is provided to justify the absence of bias.

Bias domain	Question	Coding	Scoring criteria	Decision rules
	6b. Measurement error - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
7	**7a.Analysis reporting bias: RCTs** Was the study free from selective analysis reporting?	Y, PY, PN, N, U	**a)** Authors report results corresponding to the outcomes announced in the method section (there is no outcome reporting bias). **b)** Authors report multiple analyses appropriately (e.g., by age group, sex). **c)** A pre-analysis plan or trial protocol is published and referred to or the trial was pre-registered, or the outcomes were pre-registered. **d)** Authors report appropriate analysis methods, including results of unadjusted analysis and ITT estimation, alongside any adjusted and treatment-on-the-treated/complier-average-causal-effects analysis. **e)** Analysts were blinded to treatment status.	-Score "Low risk" if all criteria are satisfied. -Score "Some concerns" if all the conditions are met except a), or if all the conditions are met but there is some element missing that could have helped understand the results better. -Score "High risk" if no pre-analysis plan or trial protocol was published or pre-registered.
	7b.Analysis reporting bias: NRS Was the study free from selective analysis reporting?	Y, PY, PN, N, U	**a)** There is no evidence that outcomes were selectively reported (e.g., results for all relevant outcomes in the methods section are reported in the results section). **b)** Authors use credible methods of analysis to address attribution given available data. **c)** A pre-analysis plan is published, especially for prospective NRS (but ideally also for retrospective studies). **d)** Requirements for specific methods of analysis: - For RDD, Researchers should analyse the change in slope and/or level using different band-widths around the threshold or functional form. The following should be pre-specified as far as possible and reported in	-Score “Low risk” if all criteria are satisfied. -Score "Some concerns" if authors combined methods and reported relevant tests (d) only for one method, or if all the criteria are met except for c) and it is a retrospective NRS. -Score "High risk" if authors use uncommon or less rigorous estimation methods such as failure to conduct multivariate analysis for outcomes equations, or if some important outcomes are subsequently omitted from the results or the significance and magnitude of important outcomes was not assessed.

Bias domain	Question	Coding	Scoring criteria	Decision rules
			sensitivity analysis: (a) selection of optimal bandwidth using existing data-driven routines; (b) selection of appropriate functional form for the relationship between assignment and outcome variables; and (c) robustness checks of other bandwidths and functional form specifications. - For PSM and covariate matching: (a) Where over 10% of participants fail to be matched, sensitivity analysis is used to re-estimate results using different matching methods (Kernel Matching techniques); (b) For matching with replacement, no single observation in the control group is matched with a large number of observations in the treatment group, and authors take into account the use of control observations multiple times against the same treatment in the standard error calculation; (c) for PSM, Rosenbaum's test suggests the results are not sensitive to the existence of hidden bias; (d) different matching methods including varying sample sizes yield the same results. - For IV models, the authors test and report the results of a Hausman test for exogeneity ($p \leq 0.05$ is required to reject the null hypothesis of exogeneity). - For Heckman selection models, the coefficient of the selectivity correction term (Rho) is significantly different from zero ($p < 0.05$).	
	7c. Analysis reporting bias - justification	Open answer	Justification for coding decision (include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

Bias domain	Question	Coding	Scoring criteria	Decision rules
8	**8a. Method used to address differences between UoA and UoR/UoT**	Open answer	Briefly describe methods used to adjust standard errors to account for correlation across units (e.g., cluster-robust standard errors reported). Unit of analysis (UoA) is the unit of observation (e.g., individual, household, community, village), unit of randomisation (UoR) is the unit of assignment to control or treatment groups (e.g., individual, household, community, village), and unit of treatment (UoT) is the level at which treatment happens (e.g., individual, household, community, village).	
	8b. Unit of analysis error: RCTs Is unit of analysis in cluster allocation addressed in standard error calculation?	Open answer	- Not applicable if it is there is no clustering in the design at household or group levels. - If UoA equals UoR, or UoA is not equal to UoR and standard errors are clustered at the UoR level, or data are collapsed to the UoR level, no adjustment is needed. - If unit of analysis errors are apparent, or insufficient information provided on the way the standard errors were calculated or what the unit of analysis is, authors should consider adjusting standard errors using variance inflation formula in sensitivity analysis.	
	8c. Unit of analysis error: NRS Are correlations between units addressed in standard error calculation?	Open answer	- Not applicable if there is no clustering in the design at household or group levels. - If UoA equals UoT, or if UoA is not equal to UoT and standard errors are clustered at the UoT level, or data are collapsed to the UoT level, no adjustment is needed. - If unit of analysis errors are apparent, or insufficient information provided on the way the standard errors were calculated or what the unit of analysis is, standard errors should be adjusted using variance inflation formula.	

Appendix B Systematic searches for internal replication studies

The information contained in this Appendix is taken from Villar and Waddington (2019). Research Papers in Economics (RePEc) database via EBSCO was searched using the following string:

> (nonexperiment* OR non-experiment* OR "non experiment*" OR quasi-experiment* OR "Quasi experiment*" OR observational OR non-random* OR nonrandom* OR "non random*" OR within-study OR "within study" OR replicat* OR "propensity score" OR PSM or discontinuity OR RDD) AND ('experiment*' OR random*)

Snowball searches were done using forwards citation tracking and bibliographic back-referencing. Drawing on this list of well-known reviews of internal replication studies below, three electronic tracking systems (Google Scholar, Web of Science and Scopus) were used to identify and screen articles that cite these reviews (forward citation tracking). Hand searches of the reference lists of all primary studies to further identify studies that had been cited in the existing literature (bibliographic back referencing) were done.

Institutional website repository searches were done using findings from a unique project extending nearly five years of systematic searching, screening, and indexing of impact evaluation across the field of international development. Further described by Sabet and Brown (2018), the 3ie Impact Evaluation Repository provides an index more than 4,000 impact evaluations populated through a project of systematic screening of more than 35 databases, search engines, and websites. It also reports descriptive information on studies key characteristics, including study design, country of origin, sectoral focus etc. This database was used to identify evidence from studies in international development that are not yet recorded in the boarder internal replication literature. All studies were screened in the repository recorded as using both a randomised and non-randomised design.

Table B1 Surveys of internal replication studies

Authors	*Title*
Bloom et al. (2002)	Can nonexperimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?
Glazerman et al. (2002)	Nonexperimental replications of social experiments: a systematic review
Glazerman et al. (2003)	Nonexperimental versus experimental estimates of earnings impacts
Cook and Wong (2008)	Empirical tests of the validity of the regression discontinuity design
Cook et al. (2008)	Three conditions under which experiments and observational studies produce comparable causal estimates
Pirog et al. (2009)	Are the alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations
Shadish and Cook (2009)	The renaissance of field experimentation in evaluating interventions
Shadish et al. (2012)	A case study about why it can be difficult to test whether propensity score analysis works in field experiments
Hansen et al. (2013)	A comparison of model-based and design-based impact evaluations of interventions in developing countries
Shadish (2013)	Propensity score analysis: promise, reality and irrational exuberance
Cook (2014)	Testing causal hypotheses using longitudinal survey data: a modest proposal for modest improvement
Fretheim (2015)*	A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation
Steiner and Wong (2016)	Assessing correspondence between experimental and non-experimental results in within-study-comparisons
Wong and Steiner (2016)	Designs of empirical evaluations of non-experimental methods in field settings
Jaciw (2016)	Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: the methodology
Wong et al. (2017)	Empirical performance of covariates in education observational studies
Chaplin et al. (2018)	The internal and external validity of the regression discontinuity design: a meta-analysis of 15 within-study-comparisons

Note: * this is a primary study covering multiple trials, which were found using systematic search methods, and re-analysed by the authors.

The web repository of a known producer of internal replications, Mathematica Policy Research Inc., was searched, as preliminary searches suggested this organization had published several internal replication studies. Therefore, Mathematica's website was screened using the search function to identify pages, documents and articles featuring the term "within-study".

The RePEc database search, conducted in August 2016, returned 3,271 records in total. Citation tracing, in August 2016, returned a further 951 records. The search of institutional repositories (Mathematica in August 2016, 3ie in January 2017) identified 307 records. Contacting authors of existing studies, and hand searches of repositories of known studies, identified 13 additional references.

Appendix C Effect size calculations

Treatment effects of continuous outcome variables were converted into the mean difference *D*, or standardised mean difference, and 95 percent confidence interval (Higgins and Green, 2011). *D* is the difference in treatment and control group means, in the units of measurement used in that study:

$$D = y_t - y_c \qquad (A1)$$

where y_t is the outcome in the treatment group and y_c the outcome in the comparison group. The standardised mean difference (*d*) measures the size of the intervention effect in each study in units of standard deviation observed in that study and is thus independent of units of measurement. The *d* statistic is the ratio of *D* to the standard deviation of the outcome, *S(y)*:

$$d = \frac{y_t - y_c}{S(y)} \qquad (A2)$$

This formula was also used for double difference estimates in which case Δy refers to the change in the outcome rather than the level:[170]

$$d = \frac{(y_{t+1} - y_t) - (y_{c+1} - y_c)}{S(y)_{t+1}} = \frac{\Delta y_t - \Delta y_c}{S(y)_{t+1}} \qquad (A3)$$

where y_t and y_{t+1} refer to pre-test and post-test measures, respectively. If studies collected pre-test and post-test outcomes data, the pooled standard deviation measured at post-test $S(y)_{t+1}$ was used.

All effect sizes were calculated so that an increase in *d* measured an improvement. For outcomes for which a negative effect was an improvement (e.g., mortality) equation 2 was multiplied by -1 (or in the case of ratio estimates, raised to the power -1).

[170] For regression-based studies the treatment mean was calculated as $y_t = y_c + b$, where *b* is the regression coefficient on the treatment dummy variable. For studies using statistical matching, the mean difference was calculated from the mean outcome levels for treatment and comparisons after matching. Where kernel matching was used, $y_c = y_t - ATET$ where *ATET* is the average treatment effect on the treated.

For the denominator, *S(y)*, the pooled standard deviation S_p was calculated:

$$S_p = \sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c - 2}} \qquad (A4)$$

where s_t and s_c are the standard deviations in treatment and comparison groups respectively, measured at post-test, and n_t and n_c their respective sample sizes.

In the case of dichotomous outcomes, many studies reported proportions, such as school attendance or enrolment, or the percentage of households using facilities. In cases where outcomes were based on proportions of events or days (e.g., disease prevalence rate), the standardised proportion difference effect size was calculated:

$$d = \frac{p_t - p_c}{S(p)} \qquad (A5)$$

where p_t is the proportion in the treatment group and p_c the proportion in the comparison group. The denominator is given by:

$$S(p) = \sqrt{p\,(1-p)} \qquad (A6)$$

where p is the weighted average of p_c and p_t:

$$p = \frac{n_c\ p_c + n_t\ p_t}{n_c + n_t} \qquad (A7)$$

and where n_c and n_t are the sample sizes of the treatment and comparison groups, respectively.

In cases where outcomes were reported in proportions of individuals, such as disease incidence, and it was necessary to estimate *d*, Cox-transformed log odds ratios were calculated (Sańchez-Meca et al., 2003):[171]

[171] Standard error of Cox-transformed *d* is given as: $se(d) = \frac{\sqrt{3}}{\pi}\sqrt{\frac{1}{n_t p_t} + \frac{1}{n_t(1-p_t)} + \frac{1}{n_c p_c} + \frac{1}{n_c(1-p_c)}}$.

$$d = \ln(OR)\,\frac{\sqrt{3}}{\pi} \qquad (A8)$$

where *OR* is the odds ratio calculated from the two-by-two frequency table:

$$OR = \frac{p_t/(1-\ p_t)}{p_c/(1-\ p_c)} \qquad (A9)$$

and 0.5 was added to all frequencies when any of them was equal to zero (Sańchez-Meca et al., 2003). Where outcomes were dichotomous, as in the case of mortality, the odds ratio was used. Where studies used regression methods, *OR* was calculated as:

$$OR = \frac{(y_c + b)/(1-(y_c+b))}{y_c/(1-\ y_c)} \qquad (A10)$$

which makes use of $p_t = y_c + b$, where y_c is the outcome mean in the control and b the treatment effect regression coefficient. In such circumstances, the standard error of the logarithm of *OR* was given by:

$$se(lnOR) = \sqrt{\frac{1}{n_t(y_c+b)} + \frac{1}{n_t(1-y_c-b)} + \frac{1}{n_c y_c} + \frac{1}{n_c(1-y_c)}} \qquad (A11)$$

Some studies reported the risk ratio, *RR*:

$$RR = \frac{p_t n_t/n_t}{p_c n_c/n_c} = \frac{p_t}{p_c} \qquad (A12)$$

with standard error of the natural logarithm of *RR* given by:

$$se(lnRR) = \sqrt{\frac{1}{n_t p_t} - \frac{1}{n_t} + \frac{1}{n_c p_c} - \frac{1}{n_c}} \qquad (A13)$$

Where treatment and control risks were available, *RR* was transformed into *OR* using:

$$OR = RR\,\frac{1-p_c}{1-\ p_t} \qquad (A14)$$

Where risks were not given, assumed risks, $\widehat{p_t}$ and $\widehat{p_c}$, equal to the median treatment and control risks from any studies in the same country measuring that outcome, were used:[172]

$$OR = RR\frac{1-\widehat{p_c}}{1-\widehat{p_c}RR} \qquad (A15)$$

Where the hazards ratio, *HR*, was given, it was converted into *RR* using the following transformation (Shor et al., 2017):

$$RR = \frac{1-e^{HR\ln(1-p_c)}}{p_c} \qquad (A16)$$

Inserting A16 into A15, it can be shown that:

$$OR = \frac{1-p_c+(p_c-1)e^{HR\ln(1-p_c)}}{-p_c e^{HR\ln(1-p_c)}} \qquad (A17)$$

The 95 percent confidence intervals used the standard error of *d*, se(d), given by:

$$se(d) = \sqrt{\frac{n_c + n_t}{n_c\ n_t} + \frac{d^2}{2(n_c+n_t)}} \qquad (A18)$$

The standard error of *D* was calculated as:[173]

$$se(D) = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}} \qquad (A19)$$

The risk difference, *RD*, and its standard error were calculated analogously:

[172] This transformation was only made in a study reporting risk ratios in Bangladesh (Hoque et al., 1999) with imputed data from Luby et al. (2018). The formula to transform RR into OR used by Clasen et al. (2015), taken from Higgins et al. (2011), is: $RR = \frac{OR}{1-\widehat{p_c}+\widehat{p_c}\,OR}$, where $\widehat{p_c}$ represents the estimated control risk.

[173] It is also common for disease incidence studies to report the incidence rate ratio, calculated as $IRR = \frac{f_t/F_t}{f_c/F_c}$, with log standard error calculated as: $se(lnIRR) = \sqrt{\frac{1}{f_t}+\frac{1}{f_c}}$, where f_t and f_c are the numbers of disease episodes in each group, and F_t and F_c the total person-time disease-free follow-up periods.

$$RD = \frac{p_t n_t}{n_t} - \frac{p_c n_c}{n_c} = p_t - p_c \qquad (A20)$$

$$se(RD) = \sqrt{\frac{p_t(1-p_t)}{n_t} + \frac{p_c(1-p_c)}{n_c}}$$

$$= \sqrt{\frac{p_t n_t (1-p_t) n_t}{{n_t}^3} + \frac{p_c n_t (1-p_c) n_t}{{n_c}^3}} \qquad (A21)$$

For studies reporting effect sizes from regression estimates on outcomes which are proportions, then:

$$d = \frac{b}{S_p} \qquad (A22)$$

where b is the effect size estimate from the regression. If the study reported p_c and p_t, *S(p)* was calculated from equation (A6).

Equation (A22) was also used for other studies reporting regression-based estimates with S_p replaced by *S(y)*, which was calculated for regression studies as (Lipsey and Wilson, 2001):

$$S_p = \sqrt{\frac{S(y)^2 * (n_t + n_c - 2) - \frac{b^2 n_t n_c}{n_t + n_c}}{n_t + n_c}} \qquad (A23)$$

Where regression studies did not report *S(y)*, the standard error *se(b)* of the test statistic for effect size estimate *b* was usually available. In such cases, the pooled standard deviation was calculated using (Borenstein et al., 2009):

$$S_p = se(b)\sqrt{\frac{n_t\, n_c}{n_t + n_c}} \qquad (A24)$$

Further, by making use of:

$$t = \frac{b}{se(b)} \qquad (A25)$$

where *t* is test statistic for the effect size estimate, it can be shown that equations (A21) and (A23) simplify to (Lipsey and Wilson, 2001):

$$d = t\sqrt{\frac{1}{n_t}+\frac{1}{n_c}} \qquad (A26)$$

In the case of equal sample sizes in treatment and control, this can be expressed as:

$$d = \frac{2t}{\sqrt{N}} \qquad (A27)$$

where $n_c = n_t = \frac{N}{2}$, and se(d) is given by:

$$se(d) = \sqrt{\frac{4}{N}+\frac{d^2}{2N}} \qquad (A28)$$

Equation (A18) was used for *se(d)* in all cases of unequal sample size, otherwise equation (A28) was used.

Where 95 percent confidence intervals were reported instead of *t* or *se(b)*, the following was used to calculate the standard error (Higgins and Green, 2011):

$$se(d) = \frac{CI^U - CI^L}{3.92} \qquad (A29)$$

where CI^L and CI^U are, respectively, the lower and upper limits of the 95 percent confidence interval. For transformations using ratio effect size estimates, such as risk or odds ratios, the natural logarithm of the ratio was used in the calculation, and exponential taken afterwards, for example:

$$se(OR) = e^{\frac{\ln(CI^U) - \ln(CI^L)}{3.92}} \qquad (A30)$$

Effect sizes and standard errors were corrected for small sample bias by applying the following correction factor (Hedges, 1981):

$$g \quad = d\left[1 - \frac{3}{4(n_t + n_c - 2) - 1}\right] \qquad (A31)$$

Where study participants were grouped into correlated clusters of observations, the following error correction formula was used to adjust standard errors (Higgins and Green, 2011; Waddington et al., 2012):

$$se(d)' = se(d)\sqrt{1 + (m-1)\rho} \qquad (A32)$$

where m is the average number of observations per cluster and ρ is the intra-cluster correlation coefficient and $1 + (m-1)\rho$ is the design effect (*Deff*). This adjustment was not applied in clustered studies where outcomes of interest were defined at the cluster level (e.g., municipality mortality rate). Usually, ρ was not reported. In studies that calculated test statistics using cluster-robust standard errors, it was possible to calculate the standard error of d using:

$$se(d') = \frac{d}{t'} \qquad (A33)$$

where t' is the test statistic for the effect size estimate b, calculated using cluster-robust methods. Where the study did not use cluster-robust methods, the value of ρ was imputed using the following approach. The variance of d, *V(d)* is calculated as:

$$V(d) = se(d)^2 \qquad (A34)$$

Inserting equation (A34) into (A32) and rearranging gives:

$$\rho = \left(1 - \frac{V(d')}{V(d)}\right)\frac{1}{m-1} \qquad (A35)$$

where *V(d')* is calculated as the square of equation (A33) and *V(d)* the square of equation (A16):

$$\rho = \left(1 - \frac{d^2}{t'^2}\right)\left(\frac{n_c\, n_t}{n_c + n_t} + \frac{2(n_c + n_t)}{d^2}\right)\frac{1}{m-1} \qquad (A36)$$

The intra-cluster correlation coefficient was imputed for studies not presenting cluster-adjusted standard errors, or where effect sizes were calculated from participant flow diagrams. The ICC taken was for diarrhoeal morbidity from Clasen et al. (2014) where equation (A33) could be

calculated, yielding 0.026. Gyorkos et al. (2013) reported ICC equal to 0.028 for school children in peri-urban Peru.

Schmidt et al. (2011) present another way to calculate *Deff*:

$$Deff = \frac{V(d')}{V(d)} \qquad (A37)$$

However, this method does not allow adjustment by studies' known numbers of clusters and observations within clusters. Hence, where m was known, equation (A34) was the preferred means of calculating ρ.

To reduce loss of information and offset perceptions of results-related choices, control groups were split by the number of treatment arms, assuming equal incidence in each group (thus affecting standard errors but not the effect size estimate). Where this was not possible, effect estimates may be combined into 'synthetic effects', by calculating an average effect, weighted by sample size, of the relevant pair-wise comparisons in these studies, and variance accounting for the correlation between correlated comparison groups from the same study. The formula for the pooled variance is given as (Borenstein et al., 2009; Waddington et al., 2009):

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} d_i\right) = \left(\frac{1}{N}\right)^2 Var\left(\sum_{i=1}^{N} d_i\right)$$
$$= \left(\frac{1}{N}\right)^2 \left(\sum_{i=1}^{N} se_i^2 + \sum_{i\neq j}^{N} r_{ij}\sqrt{se_i^2 se_j^2}\right) \qquad (A38)$$

where N is the total number of effects d_i, and r_{ij} is the correlation between effects, calculated as the mean of the correlation of treatment groups and the correlation of the control groups, and *se* the standard errors. The correlation between control arms was assumed equal to 1 where the same control group was used as comparator and 0 otherwise. The correlation between treatment arms was assumed to be 0 when combining results from different treatment groups and 1 when combining results from the same treatment groups over time. When combining results across different individuals in the same treatment group the correlation was assumed 0.5, which estimates variance at the mid-point between the two extreme cases of treating comparisons as independent (with correlation coefficient equal to 0) and most likely

underestimating the variance, or treating them as perfectly correlated (correlation coefficient of 1) and most likely overestimating the variance.

Appendix D Additional information for mortality meta-analysis

Table D1 Risk-of-bias assessments for randomised controlled trials

Study	*Outcome*	*Confounding*	*Selection bias*	*Deviations from intended intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Boisson et al. (2010)	All-cause mortality	Some concerns	Low risk	High risk	Some concerns	Low risk	Low risk	High risk
Bowen et al. (2012)	All-cause mortality	Some concerns	Low risk	Some concerns	High risk	High risk	Low risk	High risk
Bowen et al. (2012)	Diarrhoea mortality	Some concerns	Low risk	Some concerns	High risk	High risk	Low risk	High risk
Clasen et al. (2014)	All-cause mortality	Low risk	Some concerns	Low risk	Low risk	Low risk	Low risk	Some concerns
Conroy et al. (1999)	All-cause mortality	High risk	Low risk	Some concerns	High risk	Low risk	Some concerns	High risk
Crump et al. (2005)	All-cause mortality	Some concerns	Low risk	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Du Preez et al. (2011)	All-cause mortality	Some concerns	Low risk	Low risk	High risk	Low risk	Low risk	High risk
Emerson et al. (2004)	All-cause mortality	Low risk	Some concerns	Low risk	Some concerns	Low risk	Low risk	Some concerns
Ercumen et al. (2015a)	All-cause mortality	Low risk	Low risk	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Gebre et al. (2011)	All-cause mortality	Some concerns	Low risk	Some concerns	High risk	Low risk	Some concerns	High risk
Gyorkos et al. (2013)	All-cause mortality	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Jain et al. (2010)	All-cause mortality	Some concerns	Low risk	High risk	Low risk	Low risk	Low risk	High risk
Luby et al. (2004)	All-cause mortality	Low risk	Some concerns	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Luby et al. (2006)	All-cause mortality	Low risk	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Some concerns

Study	*Outcome*	*Confounding*	*Selection bias*	*Deviations from intended intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Luby et al. (2018)	All-cause mortality	Low risk	Low risk	Some concerns	Low risk	Low risk	Low risk	Some concerns
Lule et al. (2005)	All-cause mortality	High risk	Some concerns	High risk	Low risk	High risk	Some concerns	High risk
Mengistie et al. (2013)	All-cause mortality	Low risk	Low risk	Some concerns	Low risk	High risk	Low risk	Some concerns
Morris et al. (2018)	All-cause mortality	High risk	Low risk	Some concerns	Low risk	Low risk	Low risk	High risk
Nicholson et al. (2014)	All-cause mortality	Some concerns	High risk	High risk	Some concerns	Low risk	Low risk	High risk
Null et al. (2018)	All-cause mortality	Low risk	Low risk	Some concerns	High risk	Low risk	Low risk	High risk
Peletz et al. (2012)	All-cause mortality	Low risk	Low risk	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Pickering et al. (2015)	All-cause mortality	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Low risk	Some concerns
Pickering et al. (2015)	Diarrhoea mortality	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	Low risk	Some concerns
Ram et al. (2017)	All-cause mortality	Some concerns	Low risk	Low risk	Some concerns	Low risk	Low risk	Some concerns
Semenza et al. (1998)	Diarrhoeal mortality	High risk	Low risk	Some concerns	Low risk	Low risk	Some concerns	High risk

Table D2 Risk-of-bias assessments for non-randomised studies

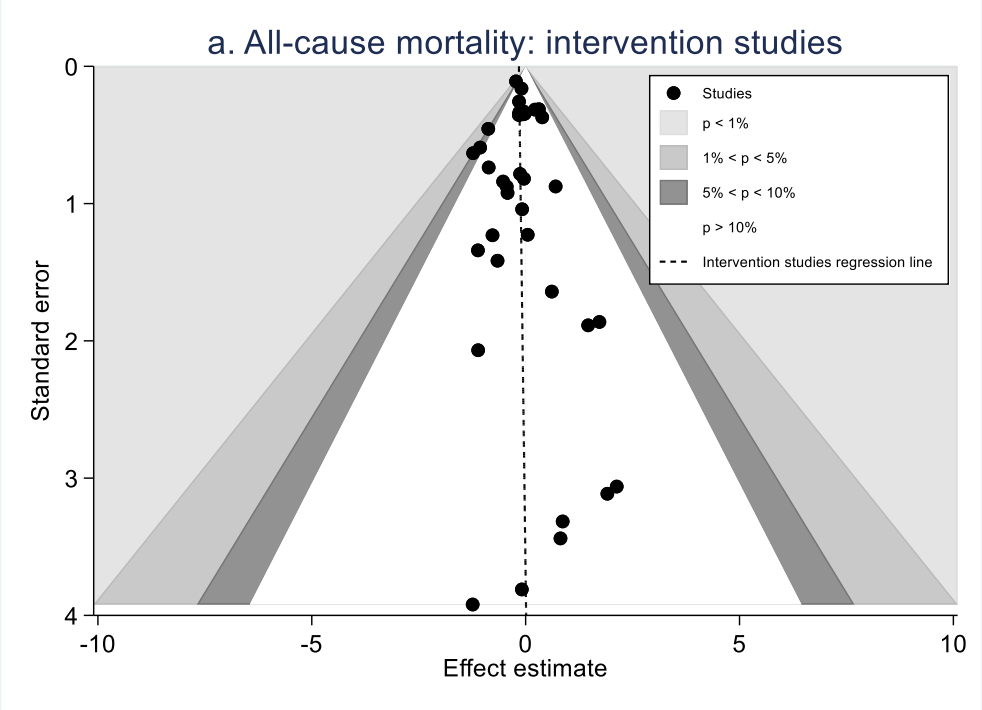
Study	*Outcome*	*Study design*	*Data source*	*Unit of analysis*	*Confounding*	*Selection bias*	*Deviation from intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Abou-Ali et al. (2010)	All-cause mortality	*Ex post* cross-section matching	National household survey	Household	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Brockerhoff (1990)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Infant, child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Brockerhoff and Derose (1996)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Infant, child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Casterline et al. (1989)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Infant, child	High risk	High risk	Some concerns	Low risk	Some concerns	Some concerns	High risk
Cole et al. (2012)	All-cause mortality	Non-randomised controlled trial	Pre-test post-test by authors	Child	High risk	High risk	High risk	High risk	High risk	Some concerns	High risk
DaVanzo and Habicht (1986)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Infant	High risk	Some concerns	Some concerns	Low risk	High risk	Some concerns	High risk
Ercumen et al. (2015b)	All-cause mortality	*Ex post* matched cohort design	Cohort survey by authors	Child	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk	High risk
Fink et al. (2011)	All-cause mortality	*Ex post* repeated cross-section OLS	National household survey	Child	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Fuentes et al. (2006)	Diarrhoea mortality	*Ex post* cross-section matching	National household survey	Infant	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Galdo and Briceño (2005)	All-cause mortality	*Ex post* repeated cross-section PSM and DD	Census data	Child	High risk	High risk	Some concerns	Some concerns	Some concerns	Some concerns	High risk

Study	*Outcome*	*Study design*	*Data source*	*Unit of analysis*	*Confounding*	*Selection bias*	*Deviation from intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Galiani et al. (2005)	All-cause mortality Infectious disease mortality	*Ex post* repeated cross-section PSM and DD	Vital registration, census data	Municipality	Some concerns	Low risk	Some concerns	Low risk	Low risk	Some concerns	Some concerns
Gamper-Rabindran et al. (2008)	Diarrhoea mortality	*Ex post* repeated cross-section FE	Census data	Municipality	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Gebretsadik and Gabreyohannes (2016)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Geruso and Spears (2017)	Diarrhoea mortality	*Ex post* cross-section OLS	National household survey	Infant	High risk	Low risk	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Granados and Sańchez (2013)	All-cause mortality Infectious disease mortality	*Ex post* repeated cross-section PSM and FE	Vital registration, census data	Municipality	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Gyimah (2002)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Infant	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Hoque et al. (1999)	Diarrhoea mortality	*Ex post* case-control	Vital registration, author survey	Child	High risk	High risk	Low risk	Some concerns	Low risk	High risk	High risk
Howlader and Bhuiyan (1999)	All-cause mortality	*Ex post* evaluation repeated cross-section	Demographic and health survey	Infant, child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Instituto Apoyo (2000)	All-cause mortality	Pipeline with group matching	Survey by authors	Community	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk

Study	*Outcome*	*Study design*	*Data source*	*Unit of analysis*	*Confounding*	*Selection bias*	*Deviation from intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Kanaiaupuni and Donato (1999)	All-cause mortality	*Ex post* evaluation repeated cross-section	Longitudinal village survey	Infant	High risk	Some concerns	Some concerns	Low risk	High risk	Some concerns	High risk
Masset and White (2003)	All-cause mortality	*Ex post* evaluation pseudo-panel	Demographic and health survey	Infant, child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Macassa et al. (2004)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Child	High risk	Some concerns	Some concerns	Low risk	Some concerns	Some concerns	High risk
Mellington and Cameron (1999)	All-cause mortality	*Ex post* evaluation cross-section	Demographic and health survey	Child	High risk	High risk	Some concerns	Low risk	Some concerns	Some concerns	High risk
Messou et al. (1997a)	Diarrhoea mortality All-cause mortality	Non-randomised controlled trial	Pre-test post-test by authors	Child	High risk	Some concerns	High risk	High risk	Some concerns	Some concerns	High risk
Rasella (2003)	All-cause mortality Diarrhoea mortality	Controlled-before-versus-after fixed effects	Vital registration data	Municipality	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Some concerns	High risk
Rhee et al. (2008)	All-cause mortality	Prospective cohort exposure design	Cohort survey by authors	Child	High risk	Some concerns	High risk	Some concerns	Low risk	Some concerns	High risk
Ryder et al. (1985)	All-cause mortality	Prospective cohort exposure design	Cohort survey by authors	Child	High risk	Some concerns	Some concerns	Low risk	Low risk	Some concerns	High risk
Reese et al. (2019)	All-cause mortality	*Ex post* matched cohort design	Cohort survey by authors	Child	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Low risk	Some concerns
Semenza et al. (1998)	Diarrhoea mortality	Prospective cohort exposure design	Cohort survey by authors	Child	High risk	Some concerns	Low risk	Low risk	Low risk	Some concerns	High risk

Study	*Outcome*	*Study design*	*Data source*	*Unit of analysis*	*Confounding*	*Selection bias*	*Deviation from intervention*	*Missing data*	*Outcome measurement*	*Reporting bias*	*Overall bias*
Victora et al. (1988)	Diarrhoea mortality	*Ex post* case-control design	Vital registration, survey by authors	Infant	High risk	Some concerns	Low risk	Some concerns	Some concerns	Some concerns	High risk

Figure D1 Funnel graphs for intervention studies



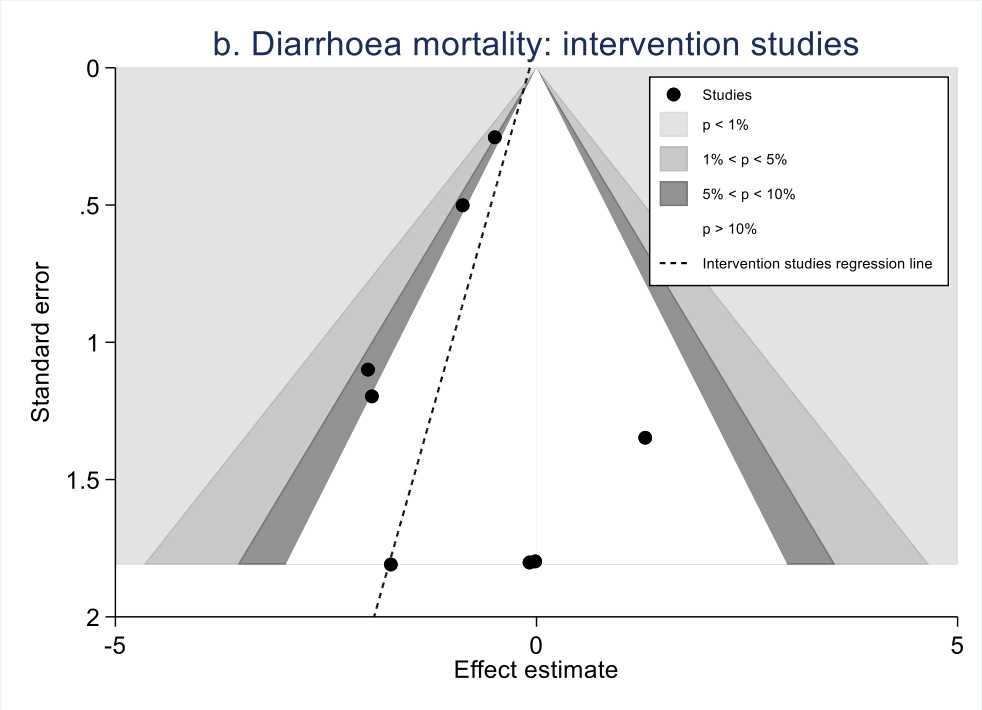


Figure D2 All-cause mortality for RCTs

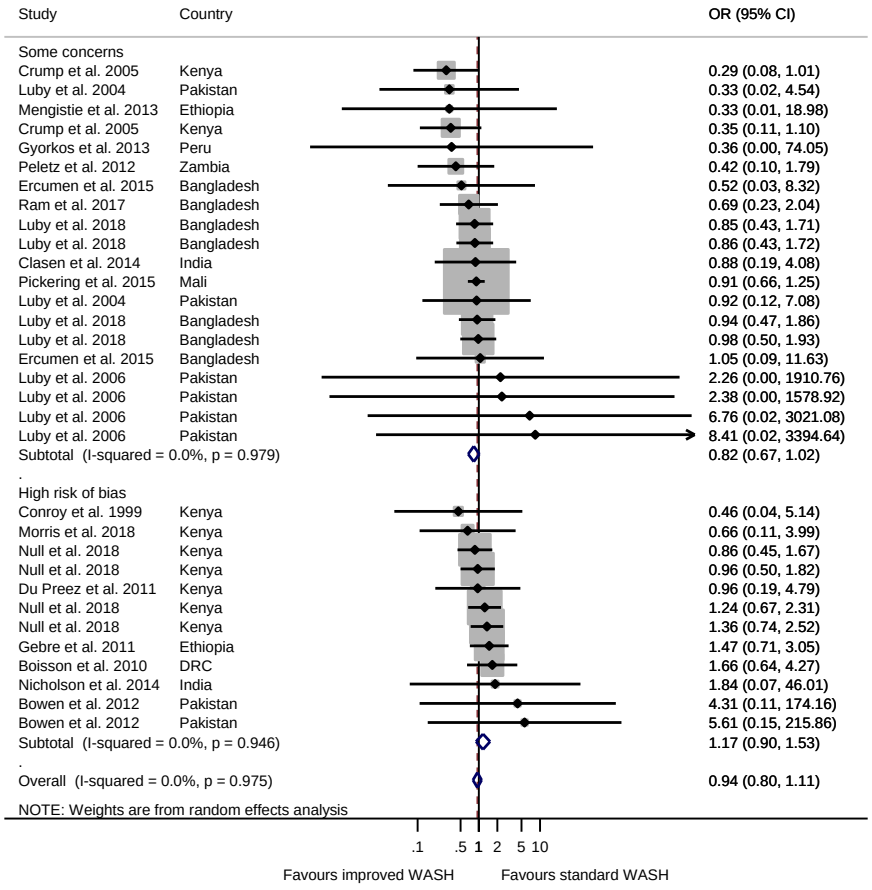


Figure D3 All-cause mortality by main WASH technology

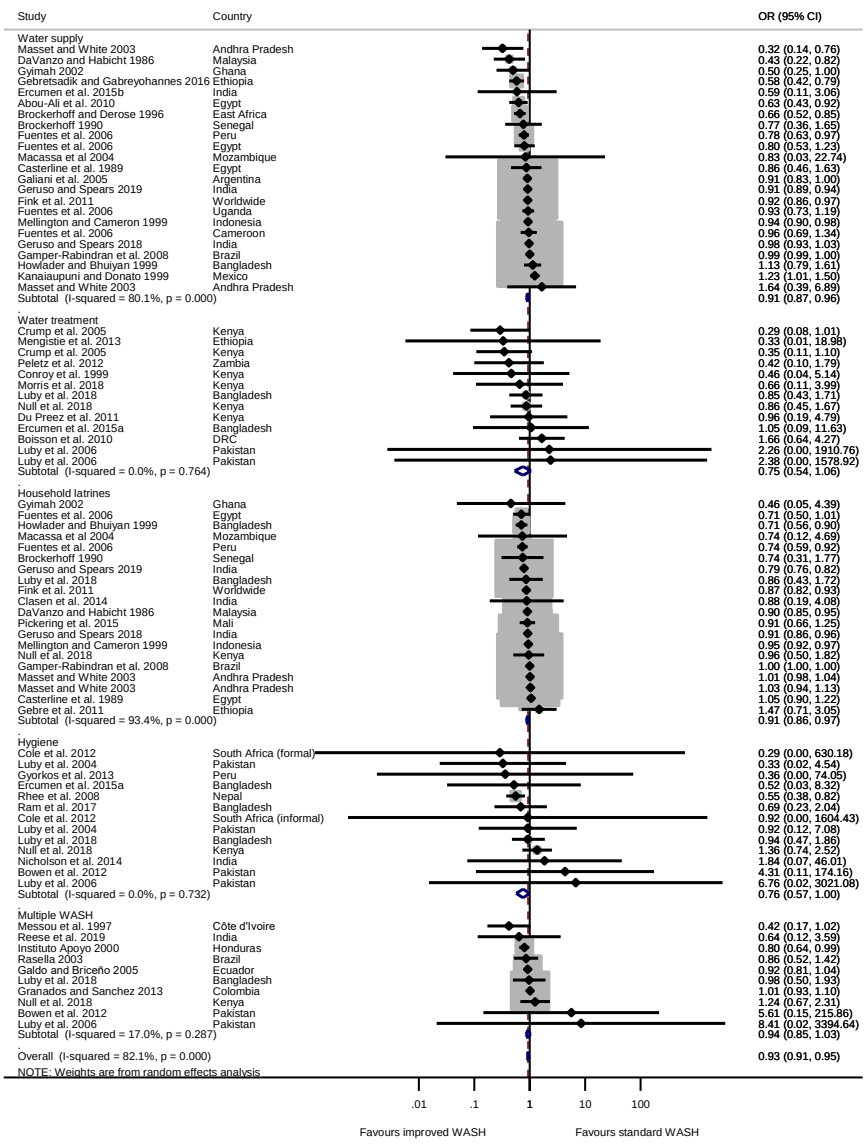
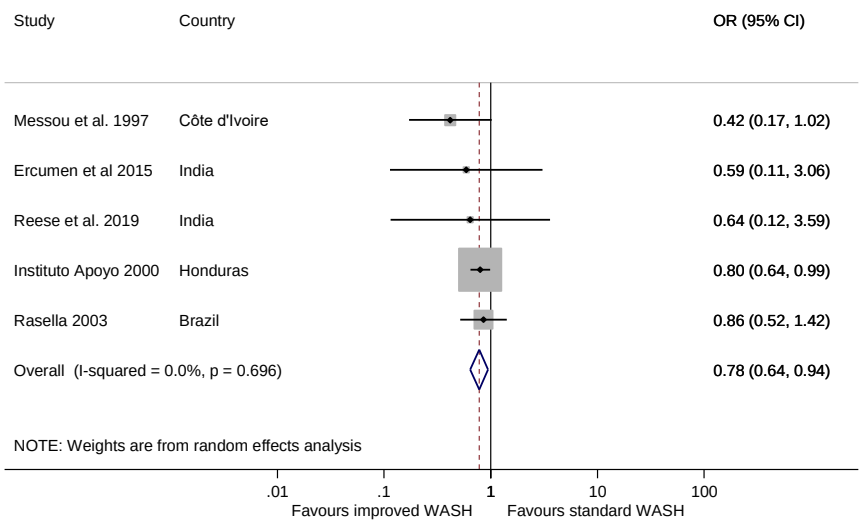
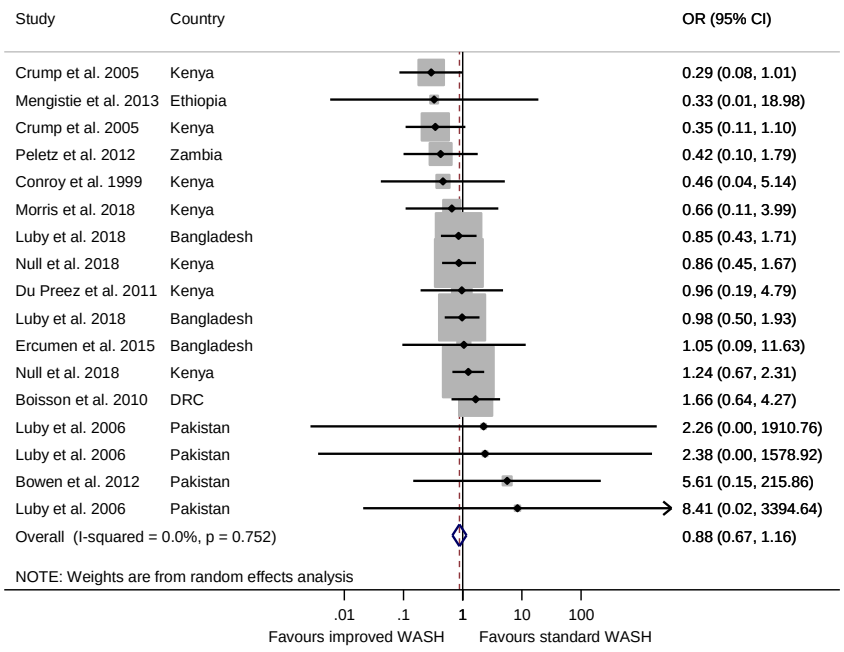


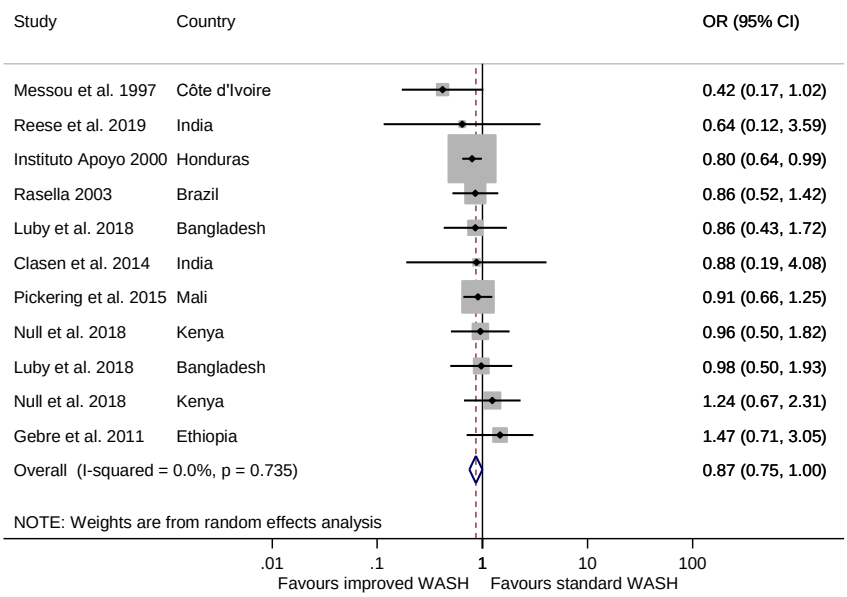
Figure D4 All-cause mortality for intervention studies



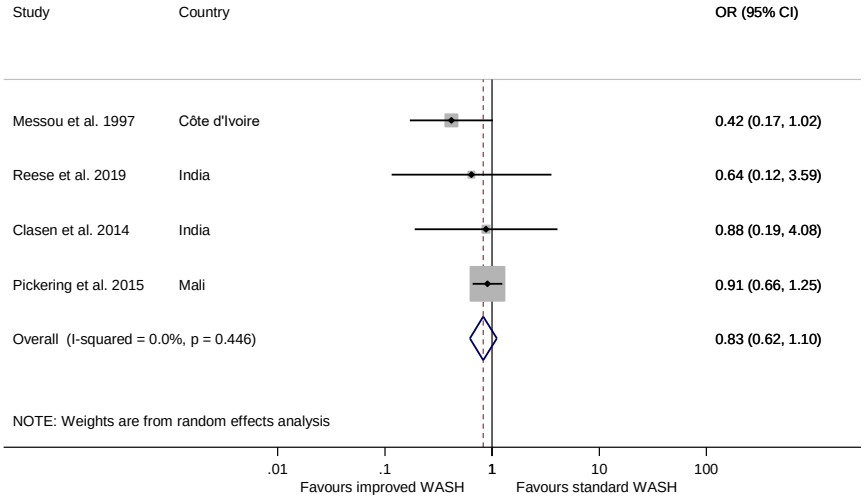
a. Water supply



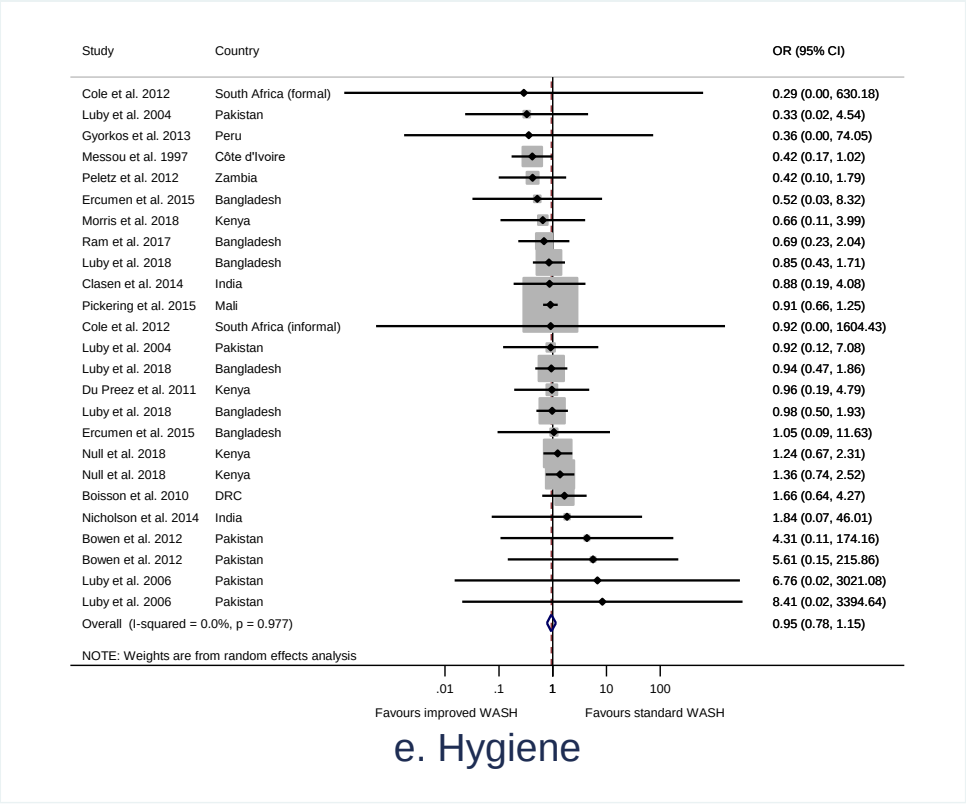
b. Water treatment



c. Household latrines



d. Latrines provided to entire community



e. Hygiene

Figure D5 Diarrhoea mortality for intervention and exposure studies

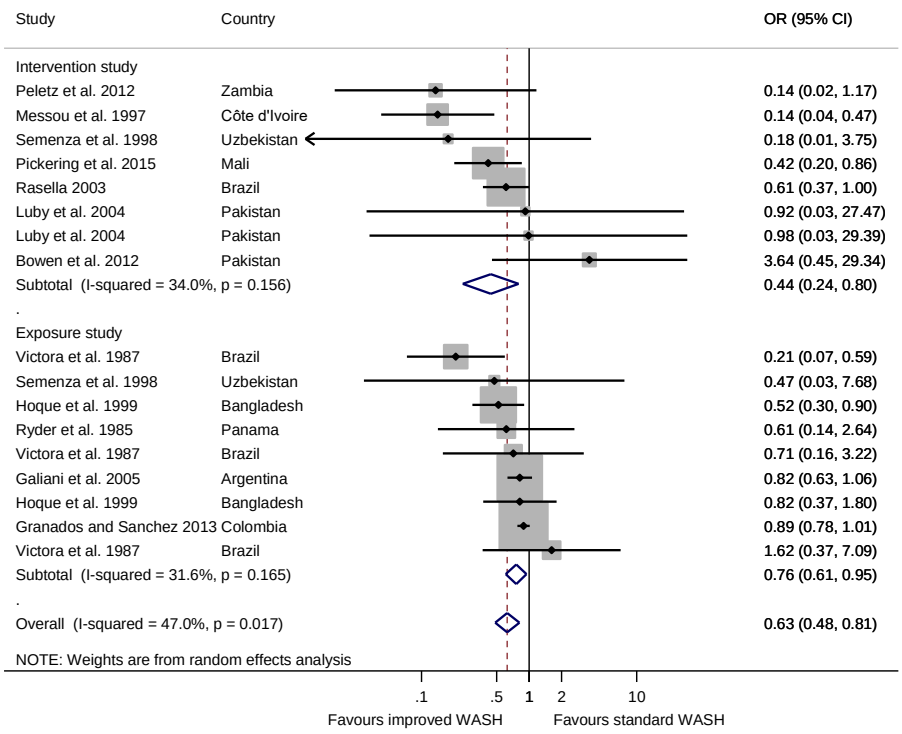


Figure D6 Diarrhoea mortality: studies with only 'some concerns'

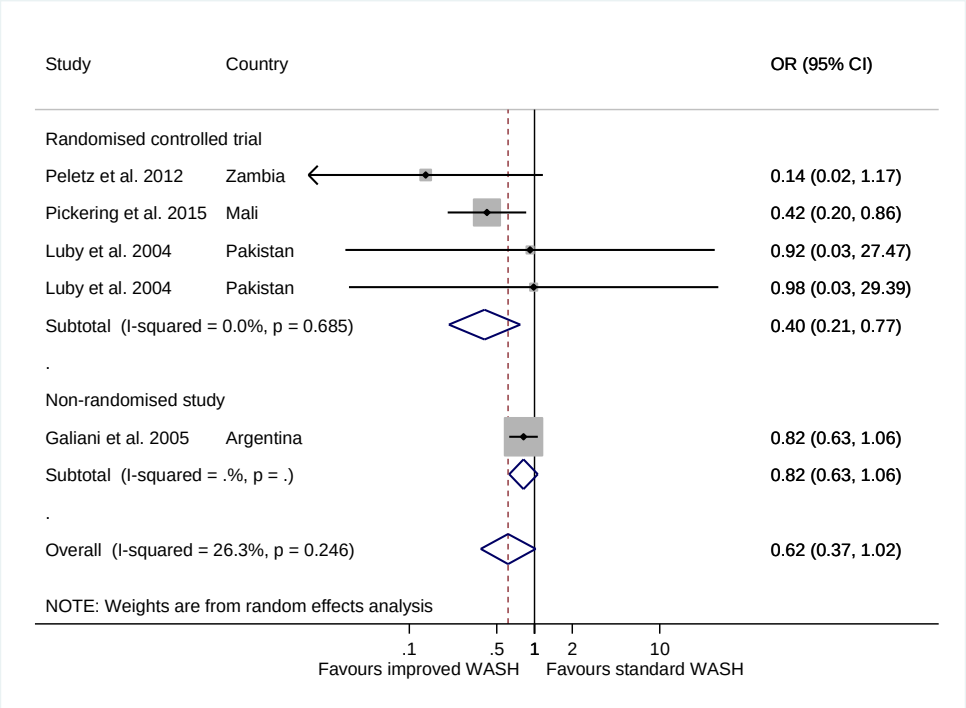


Figure D7 Mortality due to ARIs and other infectious diseases

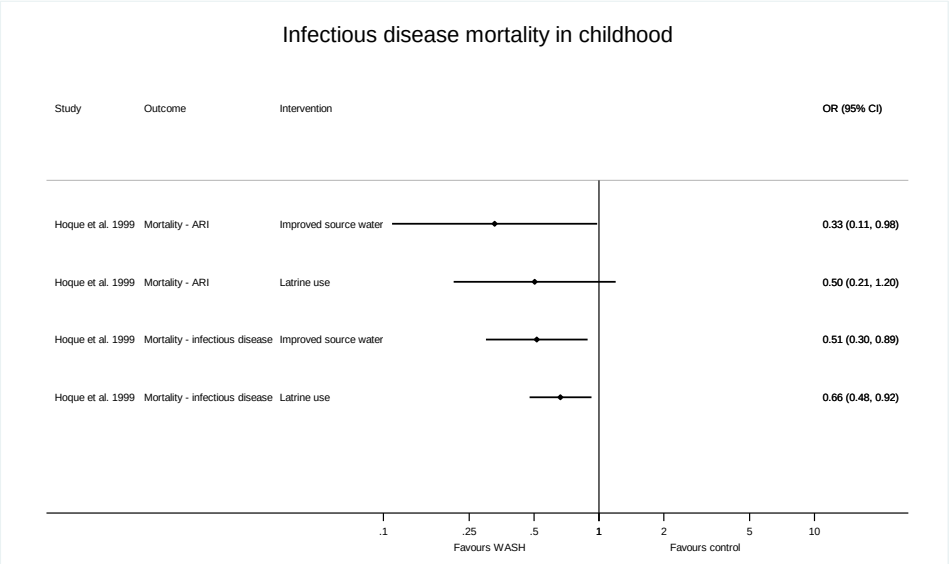
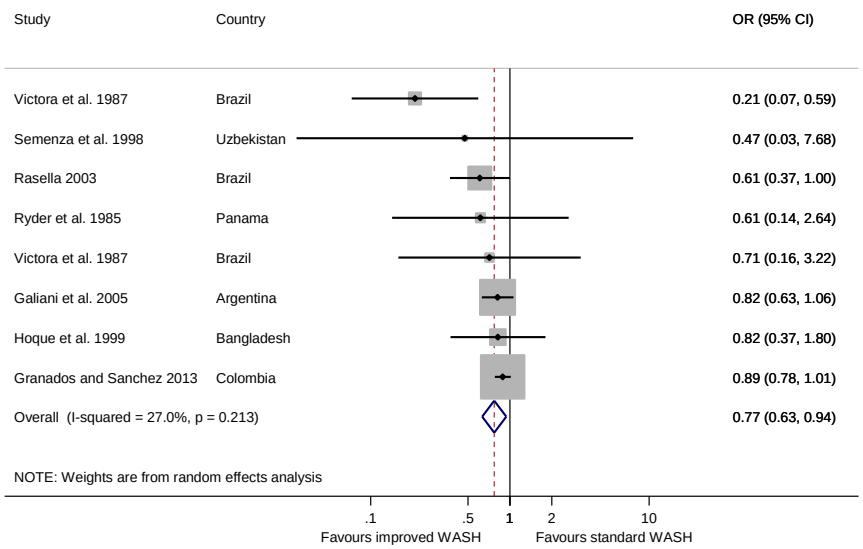
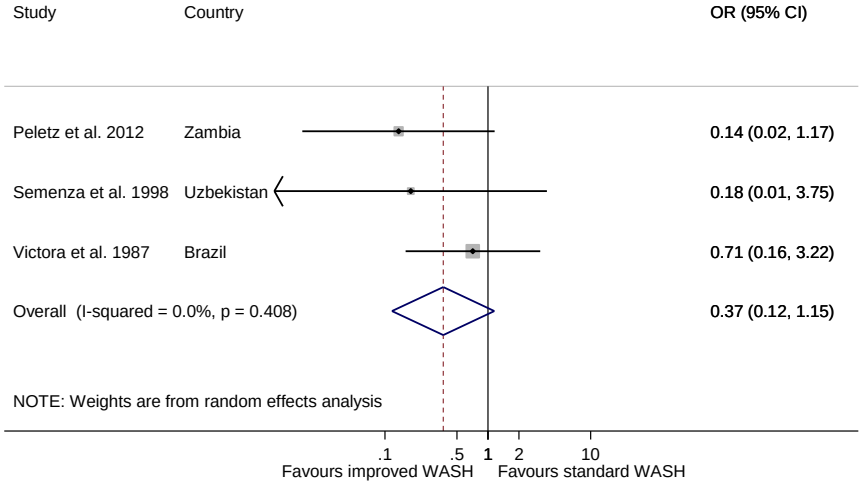


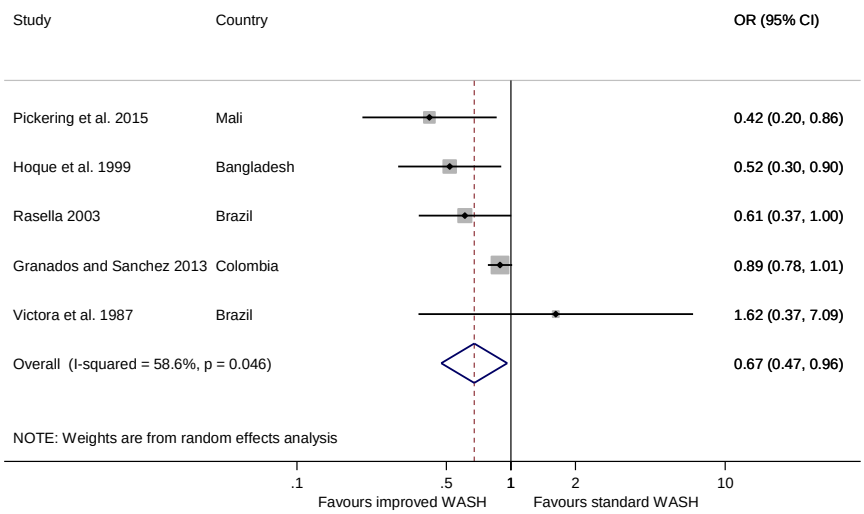
Figure D8 Diarrhoea mortality excluding Messou et al. (1997)



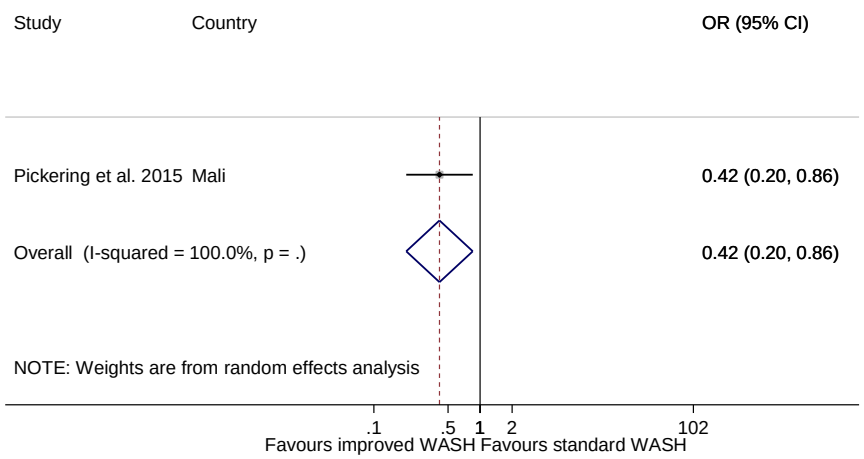
a. Water supply



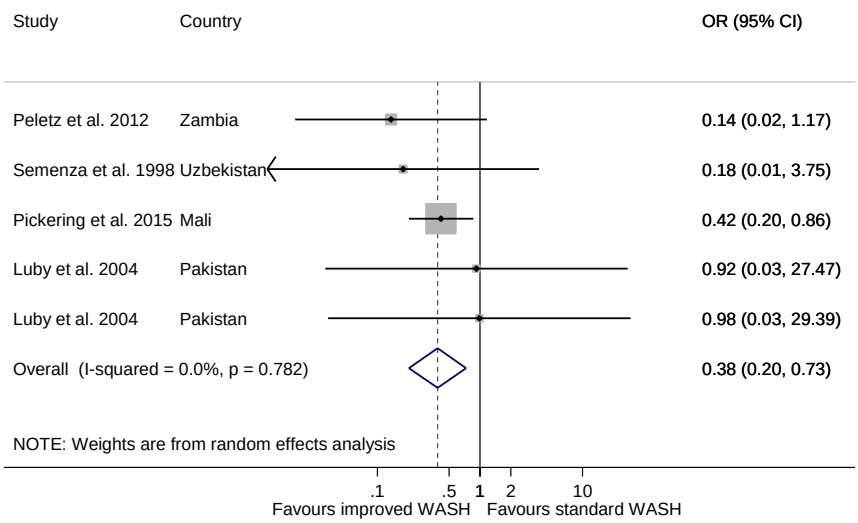
b. Water treatment



c. Household latrines

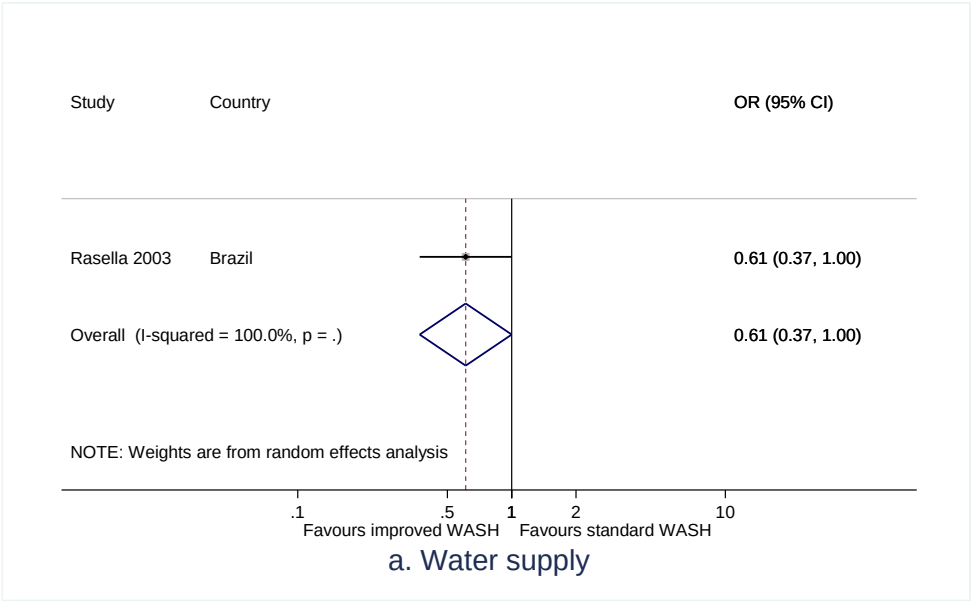


d. Latrines provided to entire community

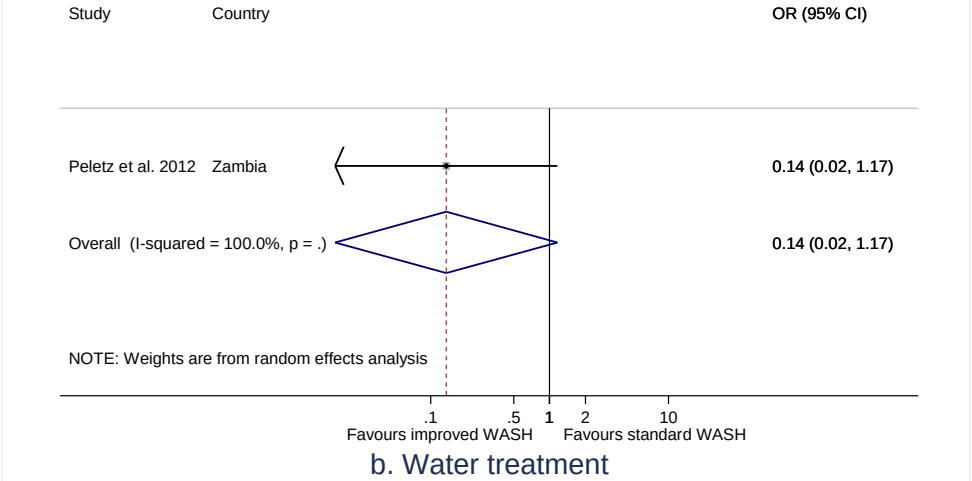

Study
Country
OR (95% CI)
Peletz et al. 2012 Zambia 0.14 (0.02, 1.17)
Semenza et al. 1998 Uzbekistan 0.18 (0.01, 3.75)
Pickering et al. 2015 Mali 0.42 (0.20, 0.86)
Luby et al. 2004 Pakistan 0.92 (0.03, 27.47)
Luby et al. 2004 Pakistan 0.98 (0.03, 29.39)
Overall (I-squared = 0.0%, p = 0.782) 0.38 (0.20, 0.73)
NOTE: Weights are from random effects analysis
.1
.5
1
2
10
Favours improved WASH
Favours standard WASH

e. Hygiene

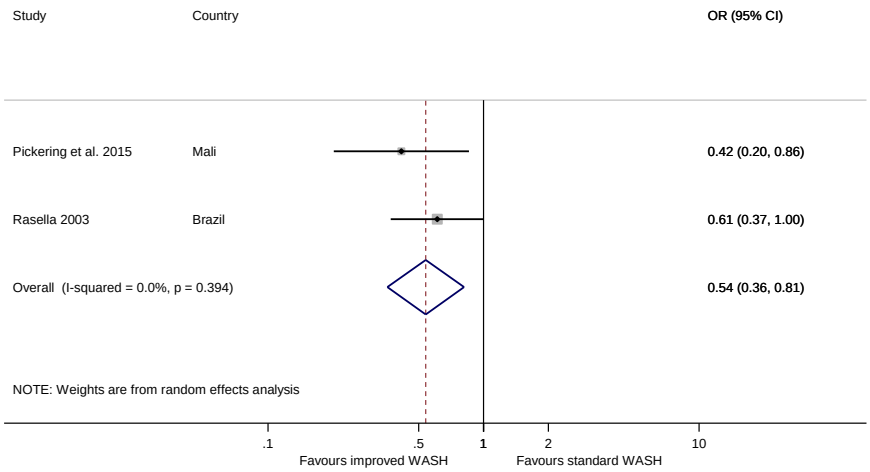
Figure D9 Diarrhoea mortality: intervention studies



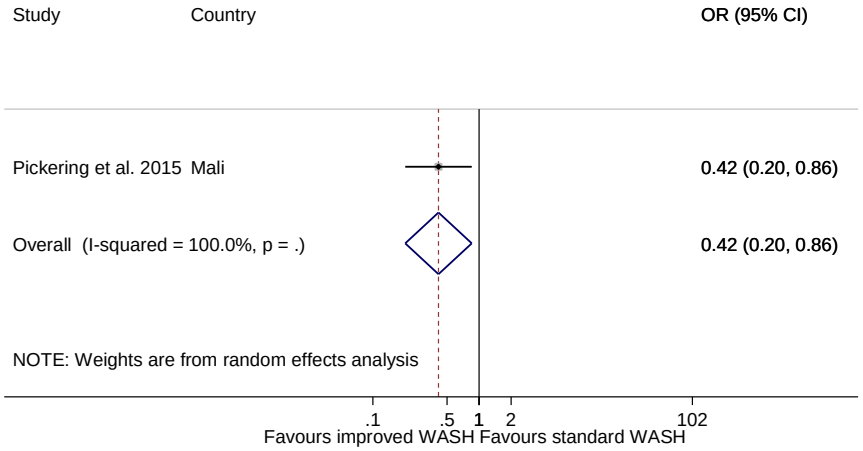
a. Water supply



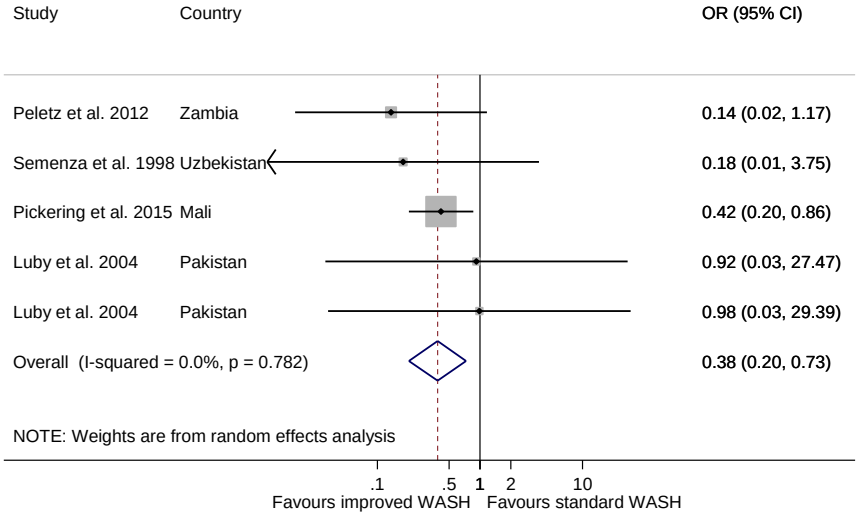
b. Water treatment



c. Household latrines



d. Latrines provided to entire community



e. Hygiene

Appendix E List of acronyms

2SLS	two-stage least squares
3ie	International Initiative for Impact Evaluation
95%CI	95 percent confidence interval
95%PI	95 percent prediction interval
ACE	Africa Centre for Evidence
AEA	American Economic Association
ANCOVA	analysis of covariance
ARI	acute respiratory infection
ATE	average treatment effect
ATET	average treatment effect on the treated
BCC	behaviour change communication
BDH	*Bono de Desarrollo Humano*
BMI	body mass index
BPL	below poverty line
CACE	complier average causal effect
CBA	controlled before-versus-after
CCT	conditional cash transfer
CDC	Centers for Disease Control and Prevention
CDD	community-driven development
CEM	coarsened exact matching
CLEAR	Centers for Learning on Evaluation and Results
CLTS	community-led total sanitation
CONSORT	Consolidated Standards of Reporting Trials
COVID-19	coronavirus disease 2019
CVM	covariate matching
DAC	Development Assistance Committee
DAG	directive acyclic graph
DALY	disability-adjusted life year
DD	double differences
Deff	design effect
DFID	Department for International Development
DHS	Demographic and Health Survey
EAP	East Asia and the Pacific
EGAP	Evidence in Governance and Politics
EPHPP	Effective Public Health Practice Project

EPOC	Effective Practice and Organisation of Care
EPPI-centre	Evidence for Policy and Practice Information and Coordinating Centre
ESI	Economics of Sanitation Initiative
FCDO	Foreign, Commonwealth and Development Office
FE	fixed effects
FFS	farmer field school
GBD	global burden of disease
GDD	geographical discontinuity design
GLAAS	Global Analysis and Assessment of Sanitation and Drinking-Water
GRADE	grading of recommendations, assessment, development and evaluations
GV	*Gram Vikas*
HAZ	height-for-age z-score
HIC	high income country
HIV	human immunodeficiency virus
ICC	intra-cluster correlation
IDCG	International Development Coordinating Group
IDRC	International Development Research Centre
IEC	information and education communication
IFAD	International Fund for Agricultural Development
IPA	Innovations for Poverty Action
IRB	institutional review board
IRC (WASH)	International Reference Centre for Water and Sanitation
ITS	interrupted time-series
ITT	intention-to-treat
IV	instrumental variables
JMP	Joint Monitoring Programme
J-PAL	Abdul Latif Jameel Poverty Action Lab
L&MICs	low- and middle-income countries
LAC	Latin America and the Caribbean
LATE	local average treatment effect
LSHTM	London School of Hygiene and Tropical Medicine
LSMS	Living Standards Measurement Survey
MCC	Millennium Challenge Corporation
MDRC	Manpower Demonstration Research Corporation
MENA	Middle East and North Africa

MHM	menstrual hygiene management
MPR	Mathematica Policy Research
MR	mortality rate
MSE	mean squared error
NGO	non-governmental organisation
NICE	National Institute for Health and Clinical Excellence
NRS	non-randomised study
NSW	National Supported Work
NTD	neglected tropical disease
ODF	open defaecation free
OED	Operations Evaluation Department
OLS	ordinary least squares
OR	odds ratio
ORS	oral rehydration salts
PAC	Pacific Access Category
PAP	pre-analysis plan
PATE	population average treatment effect
PEM	protein energy management
PHAST	participatory hygiene and sanitation transformation
PICOS	populations, intervention, comparators, outcomes and study designs
PITA	participation, inclusion, transparency and accountability
POU	point-of-use
PRAF	*Programa de Asignación Familiar*
PRISMA	preferred reporting items for systematic reviews and meta-analyses
PROGRESA	*Programa de Educación, Salud y Alimentación*
PSM	propensity score matching
RCT	randomised controlled trial
RDD	regression discontinuity design
RDiT	regression discontinuity in time
RE	random effects
ROR	relative odds ratio
RPS	*Red de Protección Social*
SANDEE	South Asian Network for Development and Environmental Economics
SHARE	Sanitation and Hygiene Applied Research for Equity
SHINE	Sanitation, Hygiene, Infant Nutrition Efficacy trial

SIGN	Scottish Intercollegiate Guidelines Network
SMD	standardised mean difference
SODIS	solar drinking water disinfection
SSIP	small-scale independent provider
SUTVA	stable unit treatment value assumption
TOT	treatment-on-the-treated
U2MR	under-2 mortality rate
UEA	University of East Anglia
UN	United Nations
UNDP	United Nations Development Program
UNICEF	United Nations Children's Fund
UoA	unit of analysis
UoR	unit of randomisation
UoT	unit of treatment
URL	*Universidad Rafael Landívar*
UV	ultraviolet
VIP	ventilated improved pit
WASH	water, sanitation and hygiene
WAZ	weight-for-age z-score
WEDC	Water Engineering and Development Centre
WHO	World Health Organization
WHZ	weight-for-height z-score
WSP	Water and Sanitation Program
WSSCC	Water Supply and Sanitation Collaborative Council
WTP	willingness-to-pay
YLD	years lived with disability
YLL	years of life lost