Things you Wanted to Know about Bias in Evaluations but Never Dared to Think

Laura Camfield, Maren Duvendack and Richard Palmer-Jones

Abstract The thrust for evidence-based policymaking has paid little attention to problems of bias. Statistical evidence is more fragile than generally understood, and false positives are all too likely given the incentives of policymakers and academic and professional evaluators. Well-known cognitive biases make bias likely for not dissimilar reasons in qualitative and mixed methods evaluations. What we term delinquent organisational isomorphism promotes purportedly scientific evaluations in inappropriate institutional contexts, intensifying motivated reasoning and avoidance of cognitive dissonance. This leads to states of denial with regard to the validity of much evaluation activity. Independent replications, revisits and restudies, together with codes of ethics that relate to professional integrity, may mitigate these problems.

1 Introduction

One of the perceived strengths of randomisation as an approach to evaluation is the way it addresses bias (Duflo, Glennerster and Kremer 2007), which has become increasingly important in an era of evidence-based policymaking and results-based management (e.g. DFID 2014). Tools such as systematic reviews posit an implicit hierarchy of evidence, based at least in part on risk of bias (Phillips *et al.* 2009). Since much of the evidence for effectiveness used by policymakers comes from evaluation (see Pritchett 2002), it is appropriate to address the prevalence of bias and reflect on possible causes. In this article we ask how, to what extent, and why different forms of bias occur in impact evaluations. First, we discuss the fragility of statistical analyses in policy contexts and the problem of false positives (the claim that an intervention produces a statistically significant effect when it does not). We then look at technical solutions to the problem of bias in the form of 'risk-of-bias' tools. We argue that one of the main weaknesses of these tools is the way they focus on internal bias, rather than considering the processes and contexts where evaluation takes place, for example the political economy of the impact evaluation marketplace and the way development evaluators are positioned within this. To balance the first section on quantitative analysis, we look next at types of bias within qualitative analysis. We then explain how delinquent organisational isomorphism can lead to the performance of evaluations that are by intention and design likely to be biased. We explore how well-intentioned evaluators manage the cognitive dissonance that occurs from the disjunction between the precept of neutral evaluation and the 'incredible' beliefs (Manski 2011) entailed by these claims. In the penultimate section we look at possible ways to address bias in evaluation. Finally, we draw some conclusions and sound warning notes for those looking for wholly technical solutions.

2 Bias in statistical analysis

It seems seldom understood how fragile even the most basic quantitative analyses can be (*cf* Manski 2007), but it is easy to demonstrate: '[W]hy Most Published Research Findings Are False', as Ioannidis (2005) puts it. Most evaluations can be posed as tests of hypotheses – specifically that an intervention had a measureable beneficial effect. In current statistical practice this will be cast as a null hypothesis that the intervention had no effect, and an alternate hypothesis that it had a sufficiently beneficial effect that outweighs the costs and adverse effects. As is well known this can result in two types of error (see Table 1) – a Type I

IDS Bulletin Volume 45 Number 6 November 2014
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

Table 1 **Evaluation findings and the real state of the world**

	State of the world	
Evaluation finding	No effect	Beneficial effect
No effect	Truth	Type II error (β) False negative Error of omission
Beneficial effect	Type I error (α) False positive Error of commission	Truth

Source Authors' own.

error when the null hypothesis is true but is rejected, and a Type II error when the alternate hypothesis is true but the null is not rejected (i.e. the alternate is rejected). The former may result in errors of commission – engaging in interventions which have not been properly shown to be effective, while the latter results in errors of omission – failing to engage in interventions which are in fact beneficial but have been rejected. Conventionally, science aimed to minimise the chance of Type I errors (implicitly minimising the chance of doing things that were not beneficial and might in fact be harmful). However, as is well known, this procedure raises the chance of making Type II errors – thereby failing to do beneficial things. The core problem is that policy science (and policy) has a pro-action bias, inducing bias in favour of positive findings.[1]

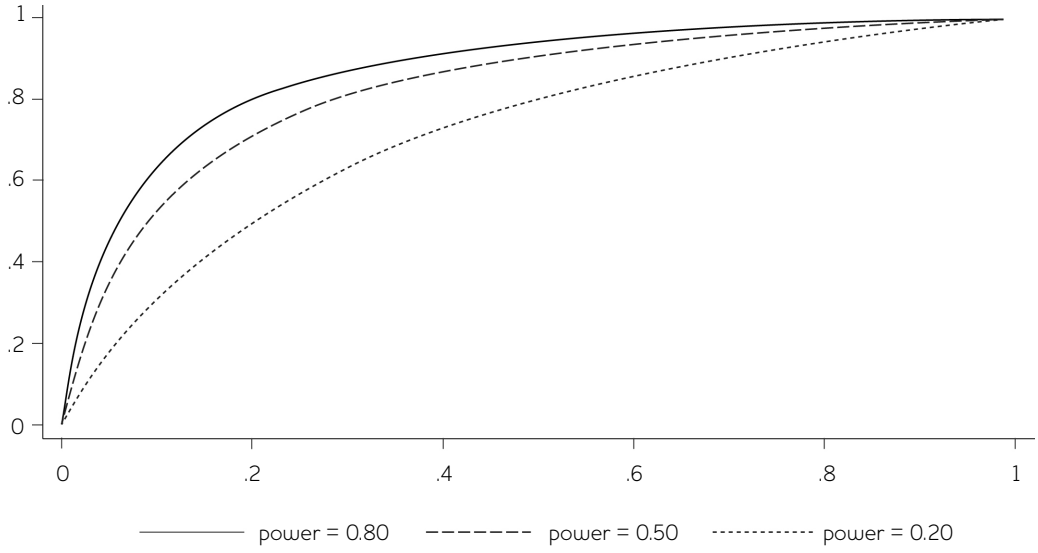
We argue next that because there are often strong pressures, for example from activist politicians, social entrepreneurs, or ambitious policy researchers, practice is in fact more oriented to avoiding Type II errors than Type I. On the academic front this arises because negative or null findings are harder to publish, while on the political or economic fronts it is because doing nothing is seldom politically or entrepreneurially rewarding. Thus it is often more important to 'find a positive result' or a big (if not particularly statistically significant) effect size than to properly discuss the limitations and uncertainties that characterise the work.[2]

As several recent authors have emphasised, modern statistical practice gives rise to unexpectedly high probabilities of false positives – rejecting the null in favour of the alternate hypotheses (Wacholder *et al.* 2004; Ioannidis 2005; Christley 2010). This arises because of multiple testing (running many analyses), data mining and model polishing (with different variables, data sets, sub-samples, estimation methods and specifications), and can be understood by considering the likely incidence of false positives. Suppose that a population of analyses (varying over the range of factors just mentioned) has a probability (Pr) of true positives[3] of a coefficient of sufficient size to warrant the intervention, then we can calculate the probability that a positive finding is true using the conventional hypothesis testing framework as follows. Generally a finding is accepted as statistically significant if only 1 in 20 positive findings are likely to occur by chance (i.e. the 95 per cent confidence level, expressed as alpha = 0.05 (1 – 0.95)). At the same time it is important that the finding should have sufficient power to fail to reject the alternate hypothesis; thus it is expected that in most cases a rejection of the null when the null is true at 95 per cent confidence should be associated with an 80 per cent chance that the null will be rejected when the alternate is true. There can be many cases when the null is false but is not rejected; the power of a test is the complement of the probability of a false negative (beta) – i.e. statistical power is 1 – beta. It can be shown that the probability of false positives (the probability that the post test result is true or the positive predictive value (PPV) of the test) is given by the formula (Maniadis, Tufano and List 2014):[4,5]

$$PPV = \frac{(1-\beta)Pr}{(1-\beta)Pr + (1-\alpha)(1-Pr)}$$

Figure 1 shows the graph of this function; clearly the probability of a true positive rises with both the frequency of true positives (or the expectation that the treatment has an effect) and the

Figure 1 **Probability of a true positive by statistical power and prior probability of positive effects**



Source Authors' own.

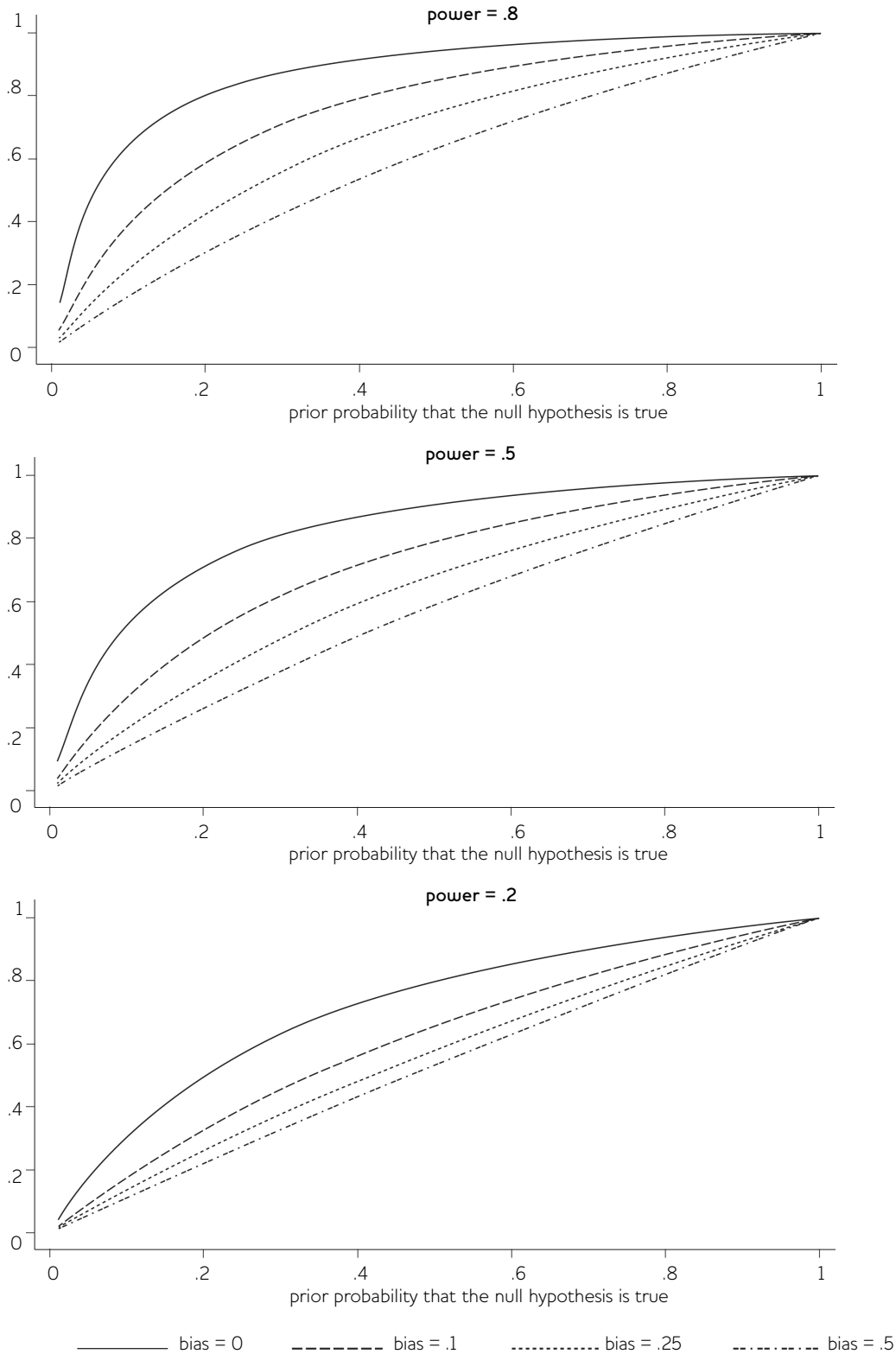
statistical power of the study. Many studies have (surprisingly) low power (Walker *et al.* 2013). Since generally we do not know the frequency of true positives for social and economic interventions the appropriate value of Pr will be subjective. Here, the politics of intervention advocacy and performance is likely to shape expectations, and will be different for different actors. Common views of effectiveness will be strongly shaped by stakeholder interests and are likely to be far from objective. Assigning a prior probability of success to a contested intervention of no more than 0.5 seems reasonable in the absence of any objective information to support the assignation of a higher value. At this prior probability (Pr = 0.5) a low powered study could give false positives in more than 20 per cent of studies. This may not seem unreasonable, although it might be hypothesised that far fewer interventions are in fact likely to be successful, until one considers the effect of study bias.

If there is bias in the evaluation study such that a fraction of u studies (or analyses within a study) produce positive results when they should not, the formula becomes:

$$PPV = \frac{(1-\beta)Pr + \beta Pru}{(1-\beta)Pr + \beta Pru + (1-\alpha+\alpha u)(1-Pr)}$$

This function is graphed for various levels of power and bias in Figure 2, which shows that if only 50 per cent of studies are biased so that they have a false positive finding, then almost 40 per cent of studies with high power will produce false positives. For studies with low power, even a modest frequency of a biased study yields a meaningful (> 40 per cent) probability of a false positive. For more unexpected or controversial interventions with say a 20 per cent chance of being successful (which are perhaps more likely to attract attention and publication) the probability of false positives can be 50 per cent for unbiased studies with low power (0.2). The prevalence of bias (u in the above function) is of course debatable, but it is surprising how common-but-questionable-practices of data analysis can readily generate a high value (see Simmons, Nelson and Simonsohn 2011 for a simulation study, and John, Loewenstein and Prelec 2012 for some empirical evidence). These phenomena are also likely causes of the 'decline effect' when scientific claims receive decreasing support over time (Schooler 2011), and explain the need for replication studies to be more highly powered than the original (Button *et al.* 2013). Low powered evaluations which produce false positives entail wasted effort in investing (more) in interventions that are not warranted, and investing less in the search for alternatives in the mistaken view that the intervention works – that is, in evaluation failure. In the following section we look at tools for managing risk of bias in the systematic review of evaluation results to see the extent to which they can address these challenges.

Figure 2 **Probability of true positive by prior probability and study power and bias**



Source Authors' own.

Table 2 **'Levels of evidence'**

Level 1a	Systematic review (with homogeneity) of randomised controlled trials (RCTs)
Level 1b	Individual RCT (with narrow confidence interval)
Level 1c	All-or-none studies
Level 2a	Systematic review (with homogeneity) of cohort studies
Level 2b	Individual cohort study (including low quality RCT; e.g. <80 per cent follow-up)
Level 2c	'Outcomes' research; ecological studies
Level 3a	Systematic reviews (with homogeneity) of case-control studies
Level 3b	Individual case-control study
Level 4	Case series (and poor quality cohort and case-control studies)
Level 5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or 'first principles'

Source Adapted from Phillips *et al.* (2009).

3 Tools for managing risk of bias within systematic reviews

Bias is understood within evaluation research as systematic deviation of results from what they should be. Most types of intervention, for example microcredit, have been evaluated in multiple studies which give differing results. For this reason systematic reviews (SRs) are recommended to synthesise diverse sources of evidence and generate reliable conclusions about the value of an intervention. In much of the SR literature freedom from bias arises through methodological rigour, reflected in the elevation of randomised controlled trials (RCTs) to the 'Gold Standard'. The RCT is placed at the top of a hierarchy of evidence (Phillips *et al.* 2009) with non-experimental methods ranked below RCTs and qualitative evidence either excluded from or ranked at the bottom; see for example Table 2.

However, much medical, natural and especially social science does not or cannot adopt RCTs, and even if they could, they are not free from bias either (the Cochrane Collaboration (2011) identifies selection, performance, attrition, detection and reporting biases as characteristic of RCTs). Alternative frameworks are therefore emerging based on typologies or matrices that judge evidence on the basis of its appropriateness to the research question, and according to its own methodology (see Petticrew and Roberts 2006). In the context of systematic reviews, it is argued that for studies to yield credible causal inference they need to be free from biases (as well as methodologically rigorous). Numerous risk-of-bias assessment tools have been developed examining the main identification assumptions underpinning the validity of quantitative studies (reviewed in Duvendack *et al.* 2012) with separate tools available to appraise the validity of qualitative evidence (summarised in Walker *et al.* 2013). These tools play important roles in the systematic review process as they ensure only high-quality studies are synthesised. Conversely they can exclude relevant evidence, which might limit our understanding of particular topics, albeit that this 'discarded' information could only be credibly included in qualitative synthesis.

Often scales are used in this process (e.g. the Maryland Scale of Scientific Methods) but many (including the Cochrane Collaboration) discourage the use of such scales (Cochrane Collaboration 2011: chapter 8.5). This led to the development of alternative approaches such as the weight of evidence (WoE) tool (EPPI-Centre 2010). WoE guides the assessment of each piece of evidence according to its methodological quality, methodological relevance or appropriateness, and topic relevance (whether the focus of the research enables the review question to be answered).

Typically these quality appraisal methods only deal with a small number of biases we are likely to encounter, typically focusing on obvious biases from weak study designs and estimation strategies. We suggest that these approaches to identifying and evaluating biases do not go far enough and that we need to also explore why

Table 3 **Types of bias: a partial list**

Empirical	Forms of cognitive bias such as sensitivity to patterns, attribution error, self-importance, halo effect
Researcher	Allegiance or experimenter bias, conservative bias, standpoint or positionality, similar person bias
Methodological	Availability bias, diplomatic bias, courtesy bias, exposure bias, bias caused through multiple mediation and distance from data generation
Contextual	Friendship bias, pro-project bias

Source Authors' own.

these biases exist and are allowed to persist. We do this by looking at the way people process information, and how information is generated and interpreted in the light of particular roles and relationships (for example, as an organisation's client, as a funded researcher, or as a member of a particular disciplinary grouping). Understanding the contexts in which bias is likely to occur may overcome a fundamental attribution fallacy – that results are due to internal (to the evaluation) factors rather than the context in which evaluations take place. We propose that research designs and methods of analysis are far more prone to bias and error than the evaluation industry likes to admit. These problems can, we argue, be attributed to the competitive context of policy analysis and the socialisation of policy analysts which leads them to minimise the gap between evaluation precepts and practices. In doing this we propose that theories of inappropriate, even delinquent, organisational isomorphism (after DiMaggio and Powell 1983) and of cognitive dissonance and motivated reasoning (after Kunda 1990) illuminate the reasons for evaluation failures, and may indicate ways to avoid them. While natural science is also characterised by practices which yield false results, poor practice in statistical and qualitative analyses thrives in evaluation environments which aspire to scientific rigour but fail to practice the most elementary procedures which characterise science, namely replication, restudy and revisit (Camfield and Palmer-Jones 2013a).

4 Bias in qualitative analysis

Bias is not restricted to quantitative evaluations. Further evidence of bias and its appearance in qualitative and mixed methods studies are discussed next. Qualitative research is subject to multiple cognitive and behavioural biases which are common to all forms of research, but perhaps more visible in qualitative. It is also particularly vulnerable to political and economic pressures, as discussed in the next section, as qualitative research within evaluation is typically associated with learning and aims to be exploratory and formative (e.g. 'developmental evaluation', Patton 2010). However, qualitative researchers do not work with a blank sheet – they have (and will be seen to have) their positionalities, which frame what they 'see' and what it is possible for them to see. Next we discuss different forms of bias (see Table 3) which we categorise as empirical, researcher, methodological and contextual.

4.1 Empirical biases

The way that evaluators engage with data is shaped by common cognitive biases. These are tendencies to think in particular ways which can lead people to make consistently poor judgements. One common bias is a tendency to see a pattern where there isn't, an example of which is the 'gambler's fallacy' (Tversky and Kahneman 1971), which causes people to over-interpret noticeable effects. Availability bias causes people to overestimate the probability of events associated with memorable or vivid occurrences such as sudden financial success. Other biases affect people's ability to assess causal relations, for example attribution bias where people see causes in internal characteristics rather than external circumstances. One example of this is where people are more likely to attribute change to specific events or actors than processes unfolding slowly over time (what Braudel (1958) called *histoire événementielle* rather than *longue durée*). Self-serving or self-importance biases may cause actors to overestimate their own contribution to social change relative to, for example domestic political processes – something that is true of agencies as well as individuals (Babcock and Loewenstein 1997). This means that respondents' narratives

cannot be treated as literal truth as they represent particular perspectives on events and processes and are further shaped by the way they were elicited.

4.2 Researcher biases

Researchers may have allegiance biases where their attachment to a particular theory causes them to discount (or ignore) other plausible explanations for similar phenomena (more generally known as 'experimenter bias' – Rosenthal 1966). Savage (2005) reanalysed archived qualitative data from Goldthorpe and Lockwood's study of car assembly workers at the Vauxhall factory in Luton (1962–3) where they tested the 'affluent worker' hypothesis by taking highly-paid workers as a 'critical case' to investigate whether everyone was becoming middle class. They concluded that this wasn't the case and Savage (2005: 39) suggests that this was because they had fitted their data into a particular typology which closed off alternative interpretations. Evaluators may also experience conservative bias where they are slow to revise their beliefs because they overweight prior evidence relative to new information.

The importance of what is often called perspective, standpoint (Harding 1987), or positionality (see also Sen 1993), illustrated by Savage's example, is something qualitative researchers are mostly aware of. Many qualitative texts include slightly awkward confessional passages (*cf* Thody 2006: 28) that locate researchers in relation to characteristics relevant to their study such as race or class. There may, however, be limited consideration of the positionality of their respondents, leading to a relative neglect of respondent biases. Other biases that occur during the research process include similar person bias, where researchers find those accounts more persuasive which come from people whom they see as similar, charismatic (*cf* the halo effect), and whom they have had personal contact with (exposure bias). As these biases are subconscious they may not be reported in reflexive accounts of the research process.

4.3 Methodological biases

A key problem in evaluation research is courtesy bias where respondents tend to tell researchers what they (are perceived to) want to hear (or what the respondents would like them to communicate; see Bavinck (2008) in relation to post-Tsunami India). Bakewell (2007: 230) describes how 'an assessment by an NGO focused on water is very likely to identify water problems that need to be addressed. Respondents will rapidly understand what the NGO's interest is and the conversation will drift towards water issues… [this] means that the data collected will reflect a particular picture that the respondents want to show in this particular context'. This may be compounded by diplomatic bias, where because the researcher is polite or timid they are reluctant to probe or challenge anomalous responses (Chambers 1983 in White and Phillips 2012: 24).

There are also processes specific to qualitative research that may increase the possibility of bias, for example the additional mediation of the data caused by interpreters, transcription and translation (Temple, Edwards and Alexander 2006), the variable quality of note-taking and the difficulty of capturing embodied knowledge (i.e. the knowledge gained by being present at the interview). Another problem is that much qualitative analysis is essentially 'tertiary' rather than secondary – done by people who neither carried out nor planned the fieldwork (*cf* White and Masset's (2007) use of data from a qualitative study of traditional beliefs related to child and maternal malnutrition to explain the lack of impact from the Bangladesh Integrated Nutrition Project, which might have resulted in bias if they had treated the data less carefully).

4.4 Contextual biases

Social and political pressures are highlighted by Boaz and Ashby (2003: 5) who note 'accusations of both conscious and unconscious bias in the peer review of research proposals, with reviewers favouring friends, protégés and those of like mind while pronouncing adversely on rivals or those with whom they are intellectually at odds'. This bias extends to the operation of steering committees and advisory groups during the research process which 'may sometimes be less to do with methodological quality than about ensuring that the project meets the requirements of the sponsor' (*ibid.*), limiting their ability to reduce bias. The bias may partly arise from relationships developing between evaluators and project staff; what White and Phillips (2012) call friendship bias (or more cynically, contract renewal bias), which then affects their independence in evaluation.[6] This

can also happen at a subconscious level if project staff become the evaluator's 'in-group' as there is a known cognitive bias towards seeing positive influences from individuals belonging to an in-group rather than an out-group (Fosterling 2001 in White and Phillips 2012). Copestake (2013) argues that the focus within impact evaluation on conventional sources of bias such as statistical sampling and selection, rather than less quantifiable ones such as pro-project bias, may be a cognitive bias in itself – towards the known over the unknown (*cf* conservative bias).

5 The politics of evaluation – cognitive dissonance and organisational isomorphism

Although most researchers are portrayed as blind to their cognitive biases, except perhaps for those who deliberately engage in fraud, we argue that biases are not 'innocent' or 'natural', and that their occurrence and effects may be understood and explained through sociological analysis of the context of evaluations. Recent calls for more evaluation in development (Savedoff, Levine and Birdsall 2005) and more learning from development evaluations (Pritchett, Woolcock and Andrews 2013) are poorly based in the history of development which is replete with the existence of earlier evaluations whose results were largely ignored (for example, the major post-Second World War projects such as the groundnuts scheme in Tanzania (Hogendorn and Scott 1981) and the Niger Valley Project (Baldwin 1957)). Social cost-benefit analysis and monitoring and evaluation were widely applied in development project planning and evaluation from the 1960s through to the present (see, for example, chapter 2 of ADB n.d.). More recently participatory appraisal and evaluation have come to characterise the field (Reitbergen-McCracken and Narayan 1998). So if we are to learn to do better evaluations and to practice evidence-based policymaking we need to understand why earlier calls for and practices of evaluation have not yielded robust practices.

First we need to recognise development evaluation failure – evaluations are conducted but they fail to bring about the sorts of sequenced improvements in interventions that have been seen in the medical and technology fields. This is partly perhaps because the model of evaluation does not fit the field of development in the same way as the field of business. In the latter, in many but not all circumstances (see Piketty 2014), competition finds out weaknesses, so that evaluation practices which do not promote competitiveness are weeded out. While some authors put this down to a lack of genuine interest in evaluation (Pritchett 2002), others have attributed it to political economy explanations that characterise evaluations in terms of principal-agent problems (Andersen and Broegaard 2012). These approaches tend to depict the actual actors in development evaluations as dumb bearers of external logics rather than consciously deliberative choosers of their actions in the evaluation field (and elsewhere). Thus we seek to characterise the evaluation field and show how it is that intelligent, educated actors come to produce evaluations that often bear tenuous relation to external reality.

Evaluation professionals have long recognised the importance of politics in evaluation and its potential for inducing bias (Datta 2011; House 1973). By its nature evaluation concerns material and financial resources, as well as their valuation, which are often disbursed by governments, or other entities not subject to market discipline.[7] Not surprisingly, organisations and people compete to appropriate them (as discussed for example in the 'rent-seeking' and 'corruption' literatures (Krueger 1974)). Organisations and individuals with major interests in interventions (as funders, suppliers, demanders, executers, resisters, evaluators, by-standers, etc.), often conceptualised as stakeholders, have multiple, diverse, and generally conflicting interests. They are likely to invest resources in the initiation, conduct, and outcomes of evaluation to ensure they better suit their needs. Medical evaluations have, according to House (2008) been captured by supplier interests, neutralising regulators and 'threaten[ing] the integrity of the evaluation field' (p.416). For this reason Datta (2011) calls for 'political toolkits' for evaluators and holds up some evaluations of contested policies as 'exemplary' (e.g. Brandon *et al.* 2010).

5.1 Evaluation implementation failure

As shown above, evaluation has been common in development in various forms dating back to the colonial period, with varying degrees of success. The recent emphasis on more and better evaluation speaks to widely perceived failures, which can be characterised as implementation

failure (Pritchett *et al.* 2013). Development organisations are induced in various ways to adopt 'best practices', generally seen as those practised in Western capitalist firms, but these do not function in the intended ways. Hence, these organisations engage in evaluation, but the evaluations are not those that are intended. These problems are not restricted to developing countries, of course, and can be found widely elsewhere. For example, among non-profits in developed countries evaluation practices have become loosely coupled to their ideal design characteristics (Weick 1974; Ebrahim 2002) so that organisations are able to control the impacts of evaluation for their own internal (management) and external (advertising) purposes.

5.2 *Delinquent institutional isomorphism*

Notwithstanding Etzioni's categorisation of complex organisations in the 1960s into those which are governed through coercive, utilitarian and normative mechanisms (Etzioni 1964, 1975), there has been a strong trend towards more homogenous forms of organisation (DiMaggio and Powell 1983) focusing on utilitarian means of eliciting compliance such as monetary payments (i.e. emphasising utilitarian means of gaining compliance to the neglect of, especially, moral means). These utilitarian forms of organisation have become increasingly characteristic of state and civil society organisations as more activities are contracted out (subject to market test, value-for-money). This trend towards homogenisation is termed organisational isomorphism; organisations increasingly appear to be governed in similar ways to those adopted in the private capitalist sector. For example, development organisations set up monitoring and evaluation departments and conduct impact evaluations with the intention of 'proving' their success and 'improving' their performance.

In the present context we draw a parallel between the rational planning procedures characteristic of business enterprises, including evaluation, and the adoption of these practices in development. Non-governmental, community-based and civil society organisations and their employees are increasingly subject to Fordist pressures manifest in evaluation procedures. However, these tendencies encounter resistance (*cf* the 'politics of evidence' movement launched by Eyben and others in 2012; see Eyben 2013).

DiMaggio and Powell (1983) introduced the terms coercive, mimetic and normative institutional isomorphisms to describe this convergence of organisational forms around the compliance characteristics of capitalist enterprises: coercive isomorphism where organisations are forced by resource dependence to adopt mandated forms; mimetic isomorphism where organisations voluntarily adopt forms of organisation which imitate those perceived as successful such as capitalist enterprises, even though they operate in different arenas and face different managerial and organisational issues and problems; and normative isomorphism where convergence occurs through professionalisation (see Section 5.3). Such organisational modelling can occur, according to DiMaggio and Powell, through transfer of personnel, or through the advice of 'consulting firms or industry trade associations' (DiMaggio and Powell 1983). Even though organisations may adopt apparently similar forms of organisation in order to improve productivity or efficacy, such transfers of organisational technology may also in part be ritualistic, for example embodying attempts at impression management or legitimisation with external or internal audiences (DiMaggio and Powell 1991: 151).[8] Normative isomorphism occurs under the influence of professionalisation of standards and values, something which is increasingly happening within evaluation. Professionalisation, according to DiMaggio and Powell, takes place as 'members of an occupation [struggle] to define the conditions and methods of their work, to control the "production of producers"' (Larson 1977: 49–52, quoted in DiMaggio and Powell 1983: 152) and to 'establish a cognitive base and legitimation for their occupational autonomy' (*op. cit.*: 152). Isomorphic pressures are intensified by organisational dependence on non-market resources (resource dependence).

Evaluation implementation failure can be explored in terms of delinquent isomorphism because coercive, mimetic, and even normative isomorphism provide no guarantee that actual practices will correspond to their putative models in institutional contexts far removed from the originals. As this is both obvious and well attested in the literatures of development (see, for example, Ferguson 1990), our interest is in what motivates the well-intentioned people who populate these delinquent forms.

5.3 *Motivated reasoning – cognitive dissonance, self-serving and other biases*

How organisations are induced to adopt forms and practices for which the institutional supports are not present should be a subject for empirical investigation. Pritchett and co-workers seem to attribute these institutional failures of isomorphic mimicry to what they term premature load-bearing due to 'the routine placement of highly unrealistic expectations on fledging systems' (2013: 1), for example a situation where a small NGO might nonetheless be pressured to adopt an elaborate M&E system. This begs the question of who places these expectations and why they are accepted. The argument of this section is that there are powerful incentives for evaluation researchers to produce positive or negative results that further their interests – mainly their careers, but also their ideologies, pet projects and so on. This results in self-serving bias which partially accounts for the biases listed in Table 3. We are not arguing that people are self-consciously, cognitively, deliberately, self-serving, rather that these dubious practices become embodied (Merleau-Ponty 1942) through processes of structuration (Giddens 1984). Thus people come to believe the truth of their statistical tests, and are genuinely puzzled, even outraged, when things go wrong (for example, failure, in the longer run, of replications). Specifically, and simplifying hugely, a concatenation of factors resulted in the dominance of Null Hypothesis Testing (NHT) in computational social (and natural) sciences which, together with the dictates of career advancement among academics – the opinion leaders in these fields – seemed to justify data mining, model polishing and HARKing (Hypothesising After the Results are Known), as well as the occasional falsification (Ioannidis 2005). These factors were included in the synthesis of Fisher's and the Neymann-Pearson methods that became the normative 'hypothetico-deductive method' of positivist, empiricist sciences (Fay 1975).

The reasons, it has been suggested, lie in the structure of rewards among empirical researchers and journal editors (Mirowski and Sklivas 1991; Feigenbaum and Levy 1993) and motivated reasoning (Kunda 1990). For various reasons the social sciences adopted NHT at the $p<0.05$ level or lower, as the requirement for consideration for publication. As many have noted, very few papers are published which report as their main finding the failure to reject the null hypothesis (Loftus 1991). Thus, according to this narrative, researchers seek to publish in high-quality journals and journals tend to publish research that reports statistically significant rejections of null hypotheses, preferably where coefficients imply effects that are meaningful to policy (i.e. the result implies a viable policy, or is substantive enough to account for a phenomenon or warrant the intervention).

Given the remarkably simple task of finding a statistically significant rejection of the null hypothesis in order to qualify for publication in prominent journals, it is not surprising that optimising individuals would seek to both conduct their analyses in order to achieve this result and to persuade themselves that what they had done was legitimate. While attribution of unreliable results to cognitive bias (confirmation bias) seems to be neutral with respect to blame, it begs the question of where these biases come from. The characterisation of the problem of false positives (and sometimes false negatives) as due to biases avoids the question of inadvertent malfeasance.

6 Strategies to address bias within evaluation

To avoid ending on too bleak a note, the final section describes some of the mechanisms we can use to reduce unacknowledged bias in impact evaluation. The first task is to increase acknowledgement of the likelihood of bias, to have more declarations of interests, broadly conceived, and to address the systemic pressures that encourage bias (Pashler and Wagenmakers 2012: 529).[9] However, we argue that the problem of bias within evaluation is systemic, embodied, and, for the most part, unconscious, inscribed by the context of socialisation and education, and exacerbated in (but not restricted to) the environment of neoliberalism. Many solutions have been proposed to address them (see Nosek, Spies and Motyl 2012) so here we rehearse some actions that address personal activities rather than structural features that reduce the unacknowledged bias in evaluation. The underlying principles for tackling bias at the level of the individual evaluation are being systematic, transparent and reflexive, and we elaborate on the implications of these next. Being systematic involves having (and publishing)[10] a clear research plan outlining the nature and sources of

data and specifying the design of instruments and protocols for fieldwork and analysis. This reduces the likelihood of researchers going on 'fishing trips' (something that is now possible also with qualitative data using Computerised Qualitative Data Analysis Software). While the more inductive nature of qualitative research makes it difficult to specify hypotheses in advance, concerns in relation to this can be allayed through transparency: giving full methodological accounts that include an account of the analysis, and archiving data to potentially enable these analyses to be 'replicated' (see Irwin 2013, for reuse of secondary qualitative research data).

Reflexivity is important in considering how the evaluator will conduct fieldwork and interviews and analyse data in a way that conveys authenticity and trustworthiness. Patton (2002: 2) argues that 'the quality of qualitative data depends to a great extent on the methodological skill, sensitivity, and integrity of the evaluator', which may be belied by the apparent simplicity of the methods (he reminds us that 'systematic and rigorous observation involves far more than just being present and looking around'). For that reason 'generating useful and credible qualitative findings through observation, interviewing, and content analysis requires discipline, knowledge, training, practice, creativity, and hard work' (*ibid.*). However, reflexivity must be demonstrated not claimed, for example by acknowledging interests, including in the field materials, and reporting your experiences, thoughts and feelings, including how your observations may have affected the observed and how you may have been affected by what you have observed.

More formal mechanisms for ensuring research quality and reducing bias include peer review and ethical codes. While peer review may identify gross examples of bias the peer reviewers rarely see the data and may well be biased themselves (Wilson *et al.* 1993; House of Commons 2011). Few if any social science journals require authors to declare their interests, let alone reviewers, and not even medical and science journals conceive interests more broadly (for example, interests in particular methods, conceptual models, or causal pathways, interventions, and so on). Grey literature and working papers may not receive the same level of scrutiny, but are still influential as, for example, Davis (2013) illustrates for Sachs and Warner's 1997 working paper on economic growth in Africa. Restudies or revisits to sites of previous studies are another way to identify gross examples of bias, or more likely, moments when the researcher's positionality and the way in which it interacted with the positionality of their participants took the research in an implausible direction (for example, the debate between Freeman and Mead over adolescent sexuality in Samoa). However, economists still seem in 'states of denial' (Cohen 2001) with regard to the value of replication in their discipline (Duvendack and Palmer-Jones 2013), although there may be some movement in this regard (Brown, Cameron and Wood 2014; Ioannidis and Doucouliagos 2013). In other work we have emphasised that research ethics extends 'beyond the subject' (Camfield and Palmer-Jones 2013b) to include the obligation to do non-trivial and beneficent research; to maintain and share accounts of research practice which affect the conclusions that can be drawn from the data; and to be clear with funders about the open-ended nature of research and researchers' additional responsibilities to society, and peers, as well as research participants. Thus, biased evaluation (research) will in the end bring evaluation (research) into disrepute (Ioannidis 2012), as of course happened with previous incarnations, for example cost-benefit analysis (Ackerman and Heinzerling 2004).

7 Conclusion

We have argued that bias in evaluation is all too easy to achieve given the normative practices of quantitative and qualitative social sciences. These practices are reinforced by the operation of cognitive biases that act to prevent psychological discomfort through the processes of motivated reasoning and are common in the evaluation context of delinquent organisational isomorphism. Unlike natural science (Merton 1942, 1973, although see Ioannidis 2012), social science is rarely self-correcting, at least in the short run. Replication is disfavoured even for entirely computational papers (Duvendack and Palmer-Jones 2013) and generally not considered possible in qualitative research, although revisits and restudies are partial equivalents (Camfield and Palmer-Jones 2013a).

Politically motivated policies are initiated and promoted by headline grabbing sound-bites and simplifying political advocacy (Cable 2004); powerful bandwagon effects come into play.

Development activists seek to expand the scale or scope of an intervention, careerist authors and editors are reluctant to write or publish articles which challenge well-established and iconic papers, and evaluation researchers may not have the time or motivation to publish. For these reasons it can take years or even decades before a more realistic assessment prevails. One might suggest that participation (and participatory research), microfinance and RCTs of social interventions are exemplars of such trajectories. In their recent incarnations starting respectively in the 1970s, 1980s and 2000s there has been discovery, sigmoid growth and more recently push-back (Cooke and Kothari 2001; Mosse 2001; Duvendack *et al.* 2011 (on microfinance); Shaffer 2011). In some cases similar interventions have been through the same cycle more than once.[11] The same should not be true of evaluation which could potentially save development from endlessly repeating its history.

Notes

1 This is, of course, not always true; Pritchett (2002) discusses logics of evaluations with different outcomes. Contesting views about this reflect in part strong motives for one or other type of finding.

2 *Vide* the perhaps apocryphal demand from President Johnson to his advisers: 'Ranges are for cattle. Give me a number' (quoted in Manski 2007).

3 Understanding can be motivated by a medical example: suppose a clinician has to interpret a symptom which has very high frequency among those with a very rare disease but also occurs at modest frequency in the population at large, and has to decide whether to prescribe a treatment that works well for the truly sick but has adverse side-effects on those with the symptom but not the disease. Failure to treat has adverse effect for the few who are afflicted, but treatment of the well also has adverse effects. The large number of false positives that are likely to occur leads to the requirement of high confidence that the patient truly has the disease.

4 Similar formulae are given in Ioannidis (2005) and Moonesinghe, Khoury and Janssens (2007), drawing on Bayes Law. In their cases Pr is substituted by the odds of true positives (R) rather than probabilities and the formulae adjusted accordingly. Note that these authors also use the more conventional alpha = 0.95 to express the confidence level of the null hypothesis, as do we.

5 Pr = prior probability that the alternate is true; alpha = 95 per cent confidence level, i.e. = 0.05 (1 – 0.95); beta = probability of a false negative; PPV = Positive Predictive Value – the post study probability that a positive finding is in fact true; u = bias.

6 The extent to which this is the norm can be seen in the hostile reaction to Mosse (2005) which project staff felt lacked the collegial and consensual approach they would expect from an evaluator (Eyben 2009).

7 We locate our necessarily truncated discussion in the arena of policy science, but acknowledge that it draws on political economy and the sociology of science.

8 The proliferation of microfinance schemes offered by civil society organisations under pressure from both clients and funders conveys the message that 'the sleepy non-profit …. Was now becoming business-minded' (Powell 1988, referred to in DiMaggio and Powell 1991), thereby appealing to the dominant neoliberal ethos of our times.

9 '[H]ypercompetitive academic climate and an incentive scheme that provides rich rewards for overselling one's work and few rewards at all for caution and circumspection' (*ibid*.: 528).

10 As now required for experimental procedures for RCTs and for systematic reviews to avoid both 'file drawer problems' (Rosenthal 1979) and Questionable Research Practices (John *et al.* 2012).

11 For example, community development (Holdcroft 1984) and RCTs of social interventions (Campbell and Russo 1998; Heckman and Smith 1995).

References

Ackerman, F. and Heinzerling, L. (2004) *Priceless: On Knowing the Price of Everything and the Value of Nothing*, New York NY: The New Press

ADB (n.d.) *Cost-Benefit Analysis for Development: A Practical Guide*, Asian Development Bank, www.adb.org/sites/default/files/cost-benefit-analysis-development.pdf (accessed 4 August 2014)

Andersen, O.W. and Broegaard, E. (2012) 'The Political Economy of Joint-Donor Evaluations', *Evaluation* 18.1: 47–59

Babcock, L. and Loewenstein, G. (1997) 'Explaining Bargaining Impasse: The Role of Self-Serving Biases', in C. Camerer, G. Loewenstein and M. Rabin (eds), *Advances in Behavioral Economics*, Princeton NJ: Princeton University Press

Bakewell, O. (2007) 'Breaching the Borders between Research and Practice: Development NGOs and Qualitative Data', in M. Smith (ed.), *Negotiating Boundaries and Borders (Studies in Qualitative Methodology, Volume 8)*, Bingley: Emerald Group Publishing

Baldwin, K.D.S. (1957) *The Niger Valley Project*, Oxford: Blackwell

Bavinck, M. (2008) 'Collective Strategies and Windfall Catches: Fisher Responses to Tsunami Relief Efforts in South India', *Transforming Cultures eJournal* 3.2: 76–92

Boaz, A. and Ashby, D. (2003) *Fit for Purpose? Assessing Research Quality for Evidence Based Policy and Practice*, London: ESRC UK Centre for Evidence Based Policy and Practice

Brandon, P.R.; Smith, N.L.; Trenholm, C. and Devaney, B. (2010) 'Evaluation Exemplar: The Critical Importance of Stakeholder Relations in a National, Experimental Abstinence Education Evaluation', *American Journal of Evaluation* 31.4: 517–31

Braudel, F. (1958) 'Histoire et Sciences Sociales: La Longue Durée', Annales, *Histoire, Sciences Sociales* 13.4: 725–53

Brown, A.N.; Cameron, D.B. and Wood, B.D.K. (2014) 'Quality Evidence for Policymaking. I'll Believe It When I See the Replication', *International Initiative for Impact Evaluation*, Replication Paper 1, Washington DC: International Initiative for Impact Evaluation (3ie)

Button, K.S.; Ioannidis, J.P.A.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.J. and Munafò, M.R. (2013) 'Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience, *Nature Reviews Neuroscience* 14.5: 365–76

Cable, V. (2004) 'Evidence and UK Politics, Does Evidence Matter?', presentation as part of an ODI Meeting Series on 'Does Evidence Matter?': 11–13, www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/206.pdf (accessed 26 August 2014)

Camfield, L. and Palmer-Jones, R.W. (2013a) 'Improving the Quality of Development Research: What Could Archiving Qualitative Data for Reanalysis and Revisiting Research Sites Contribute?', *Progress in Development Studies* 13.4: 323–38

Camfield, L. and Palmer-Jones, R.W. (2013b) 'Editorial: As Well as the Subject: Additional Dimensions in Development Research Ethics', *Progress in Development Studies* 13.4: 255–65

Campbell, D.T. and Russo, M.J. (1998) *Social Experimentation*, 1st ed., Thousand Oaks CA: Sage Publications

Christley, R.M. (2010) 'Power and Error: Increased Risk of False Positive Results in Underpowered Studies', *Open Epidemiology Journal* 3: 16–19

Cochrane Collaboration (2011) *Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0., http://handbook.cochrane.org/front_page.htm (accessed 4 August 2014)

Cohen, Stanley (2001) *States of Denial: Knowing About Atrocities and Suffering*, Cambridge: Polity Press

Cooke, Bill and Kothari, Uma (2001) *Participation: The New Tyranny?*, 1st ed., London and New York NY: Zed Books

Copestake, J. (2013) 'Credible Impact Evaluation in Complex Contexts: Confirmatory and Exploratory Approaches', Working Paper Draft, Centre for Development Studies, University of Bath, available from the author

Datta, L.-E. (2011) 'Politics and Evaluation: More than Methodology', *American Journal of Evaluation* 32.2: 273–94

Davis, Graham A. (2013) 'Replicating Sachs and Warner's Working Papers on the Resource Curse', *Journal of Development Studies* 49.12: 1615–30

DFID (2014) *How-to-Note: Assessing the Strength of Evidence*, www.gov.uk/government/uploads/system/uploads/attachment_data/file/291982/HTN-strength-evidence-march2014.pdf (accessed 4 August 2014)

DiMaggio, Paul J. and Powell, W.W. (1991) 'Introduction', in W.W. Powell and P.J. DiMaggio (eds), *The New Institutionalism in Organisational Analysis*, Chicago IL: University of Chicago Press

DiMaggio, Paul J. and Powell, W.W. (1983) 'The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields', *American Sociological Review* 48.2: 147–60

Duflo, E.; Glennerster, R. and Kremer, M. (2007) *Using Randomization in Development Economics Research: A Toolkit*, Discussion Paper 6059, London: Centre for Economic Policy Research (CEPR)

Duvendack, M. and Palmer-Jones, R. (2013) 'Replication of Quantitative Work in Development Studies: Experiences and Suggestions', *Progress in Development Studies* 13.4: 307–22

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: 456–71

Duvendack, M.; Palmer-Jones, R.; Copestake, J.G.; Hooper, L.; Loke, Y. and Rao, N. (2011) *What Is the Evidence of the Impact of Microfinance on the Well-Being of Poor People?*, London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London

Ebrahim, A. (2002) 'Implementation Struggles: the Role of Information in the Reproduction of NGO-Funder Relationships', *Non-Profit and Voluntary Sector Quarterly* 31.1: 84–114

EPPI-Centre (2010) *EPPI Centre Methods for Conducting Systematic Reviews*, Evidence for Policy and Practice Information and Co-ordinating Centre, http://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=hQBu8y4uVwI%3D&tabid=88 (accessed 4 August 2014)

Etzioni, Amitai (1975) *A Comparative Analysis of Complex Organizations: On Power, Involvement, and their Correlates*, New York NY: Free Press

Etzioni, Amitai (1964) *Modern Organizations*, Englewood Cliffs NJ: Prentice-Hall

Eyben, R. (2013) *Uncovering the Politics of Evidence and Results. A Framing Paper for Development Practitioners*, http://bigpushforward.net/wp-content/uploads/2011/01/The-politics-of-evidence-11-April-20133.pdf (accessed 4 August 2014)

Eyben, R. (2009) 'Hovering on the Threshold: Challenges and Opportunities for Critical and Reflexive Ethnographic Research in Support of International Aid Practice', in S. Hagberg and C. Widmark (eds), *Ethnographic Practice and Public Aid. Methods and Meanings in Development Cooperation*, Sweden: University of Uppsala

Fay, Brian (1975) *Social Theory and Political Practice*, London: Allen & Unwin

Feigenbaum, S. and Levy, D.M. (1993) 'The Market for (Ir)reproducible Econometrics', *Social Epistemology* 7.3: 215–32

Ferguson, J. (1990) *The Anti Politics Machine. Development, Depoliticization and Bureaucratic Power in Lesotho*, Cambridge: Cambridge University Press

Giddens, A. (1984) *The Constitution of Society: Outline of the Theory of Structuration*, Cambridge: Polity Press

Harding, S. (1987) 'Introduction: Is there a Feminist Method?', in S. Harding (ed.), *Feminism and Methodology*, Bloomington IN: University of Indiana Press

Heckman, J.J. and Smith, J.A. (1995) 'Assessing the Case for Social Experiments', *Journal of Economic Perspectives* 9.2: 85–110

Hogendorn, J. and Scott, K. (1981) 'The East-African Groundnut Scheme – Lessons of a Large-Scale Agricultural Failure', *African Economic History* 10: 81–115

Holdcroft, L. (1984) 'The Rise and Fall of Community Development, 1950–1965: A Critical Assessment', in C. Eicher and J. Staatz (eds), *Agricultural Development in the Third World*, Baltimore MD: Johns Hopkins University Press

House, E.R. (2008) 'Blowback: Consequences of Evaluation for Evaluation', *American Journal of Evaluation* 29.4: 416–26

House, E.R. (1973) *School Evaluation: The Politics and Process*, Berkeley CA: McCutchan Publishing

House of Commons (2011) *Peer Review in Scientific Publications: Eighth Report of Session 2010–12, Vol. 1: Report, Together with Formal Minutes, Oral and Written Evidence*, Vol 1, London: House of Commons Science and Technology Committee, The Stationery Office

Ioannidis, J.P.A. (2012) 'Why Science is Not Necessarily Self-Correcting', *Perspectives on Psychological Science* 7.6: 645–54

Ioannidis, J.P.A. (2005) 'Why Most Published Research Findings are False', *PLOS Medicine* 2.8: e124

Ioannidis, J.P.A and Doucouliagos, C. (2013) 'What's to Know About the Credibility of Empirical Economics?: Scientific Credibility of Economics', *Journal of Economic Surveys* 27.5: 997–1004

Irwin, S. (2013) 'Qualitative Secondary Data Analysis: Ethics, Epistemology and Context', *Progress in Development Studies* 13.4: 295–306

John, Leslie K.; Loewenstein, George and Prelec, Drazen (2012) 'Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling', *Psychological Science* 23.5: 524–32

Krueger, Anne O. (1974) 'The Political Economy of the Rent-Seeking Society', *American Economic Review* 64.3: 291–303

Kunda, Ziva (1990) 'The Case for Motivated Reasoning', *Psychological Bulletin* 108.3: 480–98

Loftus, G.R. (1991) 'On the Tyranny of Hypothesis Testing in the Social Sciences', *Contemporary Psychology* 36.2: 102–5

Maniadis, Z.; Tufano, F. and List, J.A. (2014) 'One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects,' *American Economic Review* 104.1: 277–90

Manski, C.F. (2011) 'Policy Analysis with Incredible Certitude', *Economic Journal* 121.554: F261–89

Manski, C.F. (2007) *Identification for Prediction and Decision*, Cambridge MA: Harvard

Merleau-Ponty, M. (1942) *The Structure of Behavior*, trans. Alden Fisher, Boston MA: Beacon Press, 1963; London: Methuen, 1965

Merton, R.K. (1973) *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago IL and London: University of Chicago Press

Merton, R.K. (1942) 'Science and Technology in a Democratic Order', *Journal of Legal and Political Sociology* 1: 115–26

Mirowski, P. and Sklivas, S. (1991) 'Why Econometricians Don't Replicate (Although they Do Reproduce)', *Review of Political Economy* 3: 146–63

Moonesinghe, Ramal; Khoury, Muin J. and Janssens, Cecile J.W. (2007) 'Most Published Research Findings Are False – But a Little Replication Goes a Long Way', *PLOS Medicine* 4.2: 218–21

Mosse, David (2005) *Cultivating Development: An Ethnography of Aid Policy and Practice*, London and Ann Arbor MI: Pluto Press

Mosse, D. (2001) 'People's Knowledge, Participation and Patronage', in B. Cooke and U. Kothari (eds), *Participation: The New Tyranny?*, London: Zed Press

Nosek, Brian A.; Spies, Jeffrey R. and Motyl, Matt (2012) 'Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability', *Perspectives on Psychological Science* 7.6: 615–31

Pashler, Harold and Wagenmakers, Eric-Jan (2012) 'Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?', *Perspectives on Psychological Science* 7.6: 528–30

Patton, M.Q. (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*, 1st ed., New York NY: The Guilford Press

Patton, M.Q. (2002) *Qualitative Research and Evaluation Methods*, London: Sage

Petticrew, M. and Roberts, H. (2006) *Systematic Reviews in the Social Sciences: A Practical Guide*, Oxford: Blackwell Publishing

Phillips, B.; Ball, C.; Sackett, D.; Badenoch, D.; Straus, S.; Haynes, B. and Dawes, M. (2009) *Oxford Centre for Evidence Based Medicine – Level of Evidence*, www.cebm.net/index.aspx?o=1025 (accessed 4 August 2014)

Piketty, Thomas (2014) *Capital in the Twenty-First Century*, trans. Arthur Goldhammer, 1st ed., Cambridge MA: Belknap Press

Pritchett, Lant (2002) 'It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation', *Journal of Policy Reform* 5.4: 251–69

Pritchett, Lant; Woolcock, Michael and Andrews, Matt (2013) 'Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation', *Journal of Development Studies* 49.1: 1–18

Reitbergen-McCracken, J. and Narayan, D. (1998) *Participation and Social Assessment: Tools and Techniques*, Washington DC: World Bank

Rosenthal, R. (1979) 'The "File Drawer" Problem and Tolerance for Null Results', *Psychological Bulletin* 86: 638–41

Rosenthal, R. (1966) *Experimenter Effects in Behavioral Research*, Vol. xiii, East Norwalk CT: Appleton-Century-Crofts

Savage, Mike (2005) 'Revisiting Classic Qualitative Studies', *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6.1, www.qualitative-research.net/index.php/fqs/article/view/502 (accessed 4 August 2014)

Savedoff, W.D.; Levine, R. and Birdsall N. (2005) *When Will We Ever Learn? Improving Lives Through Impact Evaluation*, Center for Global Development, Washington DC, www.cgdev.org/publication/when-will-we-ever-learn-improving-lives-through-impact-evaluation (accessed 4 August 2014)

Schooler, Jonathan (2011) 'Unpublished Results Hide the Decline Effect', *Nature News* 470.7335: 437

Shaffer, Paul (2011) 'Against Excessive Rhetoric in Impact Assessment: Overstating the Case

for Randomised Controlled Experiments', *Journal of Development Studies* 47.11: 1619–35
Sen, Amartya (1993) 'Positional Objectivity', *Philosophy and Public Affairs* 22.2: 126–45
Simmons, J.P.; Nelson, L.D. and Simonsohn, U. (2011) 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science* 22.11: 1359–66
Temple, B.; Edwards, R. and Alexander, C. (2006) 'Grasping at Context: Cross Language Qualitative Research as Secondary Qualitative Data Analysis', *FQS* 7.4, art. 10
Thody, Angela (2006) *Principles for Selecting Appropriate Writing and Presentation Styles, in Writing and Presenting Research*, London: Sage
Tversky, Amos and Kahneman, Daniel (1971) 'Belief in the Law of Small Numbers', *Psychological Bulletin* 76.2: 105–10
Wacholder, Sholom; Chanock, Stephen; Garcia-Closas, Montserrat; El Ghormli, Laure and Rothman, Nathaniel (2004) 'Assessing the Probability that a Positive Report is False: An Approach for Molecular Epidemiology Studies', *Journal of the National Cancer Institute* 96.6: 434–42
Walker, D.; Bergh, G.; Page, E. and Duvendack, M. (2013) *Adapting a Systematic Review for Social Research in International Development: A Case Study from the Child Protection Sector*, London: Overseas Development Institute
Walker, Kate; Neuburger, Jenny; Groene, Oliver; Cromwell, David A. and van der Meulen, Jan (2013) 'Public Reporting of Surgeon Outcomes: Low Numbers of Procedures Lead to False Complacency', *The Lancet* 382.9905: 1674–7
Weick, K.E. (1974) 'Education Organisations and Loosely Coupled Systems', *Administratics Science Quarterly* 2.1: 1–19
White, Howard and Masset, Edoardo (2007) 'Assessing Interventions to Improve Child Nutrition: A Theory-Based Impact Evaluation of the Bangladesh Integrated Nutrition Project', *Journal of International Development* 19.5: 627–52
White, Howard and Phillips, Daniel (2012) *Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework*, New Delhi: International Initiative for Impact Evaluation
Wilson, Timothy D.; DePaulo, Bella M.; Mook, Douglas G. and Klaaren, Kristen J. (1993) 'Scientists' Evaluations of Research: The Biasing Effects of the Importance of the Topic', *Psychological Science* 4.5: 322–5