**COMMENTARY ARTICLE**

# Complex systems, explanation and policy: implications of the crisis of replication for public health research

Robert L Grant[a]* and Rick Hood[b]

[a] Centre for Health and Social Care Research, Kingston University & St George's, University of London, London, UK. Email: robert.grant@sgul.kingston.ac.uk Telephone: +44 020 8725 0142
[b] Department of Social Work and Social Care, Kingston University & St George's, University of London, Kingston upon Thames, UK. Email: rick.hood@sgul.kingston.ac.uk Telephone: +44 020 8417 6572
* Corresponding author.

## Abstract

Much public health research considers interventions that influence and are influenced by both individuals' health and the society around them; this can be described as a complex system. We consider the role of explanation alongside statistical inference as crucial to obtaining credible and useful insights, particularly in light of concerns about a 'crisis of replication', and reflect on the difficulty in researching complex systems. In this paper, we make a connection with Lipton's philosophy of inference to the best explanation, and offer some tentative practical recommendations, extending suggestions from different fields. An extended hypothetical example is given on introducing 'sugar tax' in England.

Keywords: complexity, crisis of replication, inference to the best explanation, methodology, policy, public health, sugar tax

## Introduction

### *Complex systems*

This paper will explore the implications of complexity for the relationship between research and policy in the English public health sector. Our focus is on the interface between scientific explanation and policy prescription, drawing on the hypothetical example of a sugar tax policy as well as recent debates about statistical significance and the crisis of replication in research findings. To begin, we outline why theories of complex systems are relevant in policy and research.

Published in Critical Public Health (2017); 27(5): 525–32.
(doi: 10.1080/09581596.2017.1282603)

A key feature of a complex system is the involvement of intelligent agents, whose behaviour adapts in response to interventions, and also potentially to being observed by a researcher (Cilliers, 2005). These will typically be policy-makers, managers, professionals and service users, but increasingly may also be machine learning / artificial intelligence programs that monitor and respond to data. Complex systems are characterised by a high degree of interconnectedness between agents and their social practices (Byrne, 1998; Harvey & Reed, 1996; Hood, 2014; Stacey, 2007). New global patterns of behaviour emerge from local interactions in a way that is sensitive to but not predictable from the recent past, so it is dynamic (the rules of interaction can change) and non-linear (a small input can cause a large change) (Byrne, 1998; Sayer, 1992).

In the field of healthcare, these ideas have gained increasing prominence in the evaluation of what have been termed 'complex interventions' (Greenwood-Lee, Hawe, Nettel-Aguirre, Shiell, & Marshall, 2016, Moore et al., 2015; Campbell et al., 2007; Craig et al., 2008). According to the Medical Research Council (MRC)'s guidance on evaluating complex interventions (Craig et al., 2008), complex interventions are characterised by:

- Many interacting components within the intervention
- Requirement for those delivering or receiving the intervention to behave in a range of different ways
- Targeting of multiple groups or organisational levels
- Number and variability of outcomes
- Need for a degree of flexibility or tailoring

These features require an understanding of causal processes as well as outcomes, and acknowledge that complex interventions may work best if tailored to local circumstances. The guidance also noted the problem of replication, evidence synthesis, and wider implementation, suggesting that detailed descriptions of interventions might help to transfer them to different settings.

The MRC guidance has been influential and helpful in generating structured approaches to evaluating complex healthcare interventions, for example, Wight, Wimbush, Jepson, and Doi (2016). They have the virtue of establishing a theory of causal processes – i.e. how the intervention works – alongside the testing of outcomes through various forms of (preferably experimental) research designs, which has previously been lacking (Lewin, Glenton, & Oxman, 2009). What is perhaps less evident in this literature is a critique of the statistical models used to generate findings and how these findings are then interpreted by the wider healthcare and policy community. We believe such a critique is necessary because, in a complex system, standard statistical approaches will interpret the dynamics and non-linearity as unexplained variation, which appears to grow as more data are collected and the dynamics of the system take effect.

Typically, a statistical model will use all the available data to calculate the point estimate as well as the unexplained variation around this point. The model relies on nothing unexpected happening to disrupt the estimate, and variation is assumed to be independent of the data and is simply seen as 'noise' rather than being of intrinsic interest. These assumptions can prove misleading when they ignore contextual factors.  .The appearance of clear and compelling inferences can be achieved by reducing the data in terms of location, timescale or perspectives, for example by pooling data from different localities before and after a national intervention. Unfortunately, complex system dynamics mean that such clear findings are unlikely to persist, , and so the predicted effect of any intervention designed around them will fail to materialise. Manski called this problem 'incredible certitude' (Manski, 2013). Alternatively, strong assumptions can be imposed on the data, simplifying the analysis even if the data are not reduced.

***Inference to the best explanation***

'Inference to the best explanation' is a philosophical theory of how we make inductive inferences that emphasises the importance of explanation alongside (and distinct from) statistical inference (Lipton, 2004). Statistical inference aims to make statements about the population based on a sample, while explanation in this sense is a convincing account of causation that can aid predictions of future behaviour. Explanation is generally formed by researchers thinking about the results and is not determined objectively by the inference. Much of our notion of good practice as scientists comes from the hypothetico-deductive model from the physical sciences and experimental medical research, where a statistical inference from sample to population can reasonably be expected to reveal a pattern that will persist over time and settings. In such fields, the explanation is relatively unimportant but in complex systems it becomes more important than the statistical inference.

In recent years, much has been written about a 'crisis of replication': statistically significant patterns disappear when efforts are made to replicate the study (Baker, 2016). Often, this indicates being led by the data to the strongest signal and then regarding that as if it was a pre-specified, isolated, precise hypothesis. The most obvious form of this is running multiple statistical tests on the same data without appropriate adjustment, which is called 'p-hacking', 'data dredging' or 'fishing'. This is often done without malice but through a lack of awareness. It also happens where broad 'scientific' or 'substantive' hypotheses permit many statistical hypotheses. To take an actual example, researchers considered whether a woman's choice of clothing on a particular day was in some way linked to menstrual cycle (a broad substantive hypothesis), and presented a statistically significant difference in the proportion of pink or red shirts and tops in the 5 days around expected ovulation (a precise statistical hypothesis, one of many) (Gelman & Loken, 2013). This is a manifestation of the inference/explanation divide, as a statistical test for one explanatory model does not distinguish it from other competing models (Mayo & Cox, 2006). Platt's recommendation

for ‘strong inference’ involved an inductive iterative process, with each iteration containing a deductive experiment testing a hypothesis, and he argued against drawing conclusions from one test (Platt, 1964). Without a potential explanation, one cannot define a precise hypothesis, and may select the most interesting pattern without coming under suspicion of multiple testing; indeed, only one statistical test might be conducted, or none at all. This problem has been called ‘researcher degrees of freedom’ or ‘the garden of forking paths’ (Gelman & Loken, 2013). We prefer to emphasise the parallel with multiple testing by calling them overt multiplicity and cryptic multiplicity (see also Simmons, Nelson, & Simonsohn, 2011)).

Over-simplified analyses in complex systems have been subject to numerous criticisms over the years. By attempting to reduce complexity to a set of technical challenges, the social quasi-experiment takes place in an illusory world of control and predictability that may be reassuring but ultimately misleads policy makers (Hood, 2014). More recent approaches to evaluation have therefore highlighted the importance of programme theory (Pawson, 2013; Donaldson, 2007). Despite this attention, it is not at all clear how studies should best report their definitions and use of complexity (Thompson, Fazio, Kustra, Patrick, & Stanley, 2016). Pre-specified analysis plans have been crucial in improving the conduct of clinical trials, but when applied to complex systems do not provide details of analytical simplifications, so there is considerable scope for overt and cryptic multiplicity.

**The near future**

Readers are invited to consider a hypothetical scenario:

> ‘The British government is under pressure from business lobbyists to step back from the promised implementation of a ‘sugar tax’ on the sale of soft drinks. The new Minister for Health announces that the policy will be phased in across different regions “to allow an independent and scientific evaluation”. She tells reporters that “We are absolutely committed to doing the right thing, and doing it in the right way. Our policies are based on scientific evidence rather than opinion and hearsay. So if the science says there are no benefits to public health, we will stop, and if it says the tax delivers what it is supposed to, we will go ahead.” A large sum of money is fast-tracked to a group of public health researchers and they set about planning the work.
>
> ‘The researchers negotiate a ‘natural experiment’, in which the tax is imposed in some jurisdictions and not in others, with similar demographic characteristics in each population. Desired outcomes include sales of sugary drinks and foods, as well as health outcomes such as rates of childhood obesity and glycaemic control in people with diabetes. While the government wants results delivered within one year, the researchers do not believe this is long enough to measure the effect. A compromise is reached, with interim

outcomes (obesity rates from school nurses) reported at one year and limited funding provided for a three-year follow up using inexpensive data from existing national health surveys and voluntarily contributed data from supermarkets.

‘The research attracts attention in the media, not all of it positive. Advocates of the tax complain that evaluation is causing unnecessary delay that will ‘cost lives’, while opponents argue that it is ‘nanny state’ interference, a waste of time and money, and ethically unsound. Meanwhile, it turns out that only one supermarket chain is willing to release sales data, which will affect the validity of longer-term outcome measures.

‘After one year, the interim report is sent confidentially to the Department of Health. They do not publish it in full but rather a simple summary showing a statistically significant reduction in obesity rates. Soon it is announced that the sugar tax will therefore be implemented nationwide.

‘Concerns and caveats emerge later, but get little media coverage. The supermarket data is even more limited than expected, and shows no significant change in sales. The supermarket’s chief executive retired towards the end of the ‘natural experiment’ and gave interviews making it clear that he viewed opening up their data as a legacy of his career, “giving something back for the next generation”. Other members of the board were less keen and so, on his departure, the scope of the data was greatly curtailed and interviews arranged with staff members were cancelled, for fear of adverse publicity. Anecdotally, the researchers learn that store managers, reticent to have falling sales visible to their management, ran promotions and placed drinks more visibly within the stores to counteract it. Meanwhile, local government inspectors, charged with policing the imposition of the tax on imported drinks, had no additional funding for this, despite stories emerging of independent shops and online traders selling bulk drinks tax-free from other regions.

‘The school nurse data on obesity rates also had caveats. Although a significant reduction was observed, there were many competing explanations:

- the tax led parents to switch to healthier alternatives when shopping
- the publicity surrounding this controversial initiative, not the tax, was responsible for parents adopting healthier diets
- public health professionals saw the sugar tax pilot as an opportunity to run publicity campaigns around healthy lifestyles, and these were responsible
- the economy was improving at the time and parents were coincidentally able to afford healthier diets

- regression to the mean: the region was chosen because the previous year's obesity statistics were poor, and this happened to be a natural fluctuation which corrected itself the following year
- school nurses actively campaigned for the pilot to come to their region by weighing more children who might be around the threshold for obesity, and this campaign ended when the tax was imposed (and perhaps some unconscious bias came into weighing in the following year; they want and expect to see their efforts leading to healthier children)
- parents who felt unfairly victimised by the tax and the surrounding publicity did not cooperate with the visiting school nurses, and their children were more likely to be obese
- the change reflects a long-term trend toward healthier lifestyles nationally, not any effect of the tax itself.

Some of these problems would be resolved in an ideal world with plenty of data and a sophisticated analysis, but even the ideal hypothetico-deductive study will not help to choose an explanation. This story is designed as a critique of the way in which research comes to provide evidence for policy implementation in public health. Although fictional, it is hardly far-fetched: at the time of writing, a tax on drinks with high added sugar levels is indeed proposed for implementation in England in 2018, and a Mexican study is one of the key pieces of evidence adduced in support of the policy (Colchero, Popkin, Rivera, & Ng, 2016). There are reasons to be sceptical of the government's claim to be basing policy on scientific evidence. In the run-up to the policy's official announcement in 2016, the Chief Medical Officer for England, Dame Sally Davies, asserted in a radio interview that "we *know* it works because there is rigorous evidence from Mexico" [our emphasis] (BBC, 2015). This leaves no doubt that the Mexican findings are not only generalizable to an entirely different country and healthcare system, but constitute 'knowledge', which brings it into the realm of Manski's 'incredible certitude'.

Colchero and colleagues report one year after implementation of the tax in Mexico (Colchero et al., 2016). Their evaluation contrasts observed beverage sales with a projection based on two years of data prior to the tax. The study appears to have been conducted to a high standard, and the paper does not make unsupported claims, though two major limitations are appropriately highlighted by the authors. First, two years is insufficient data to model a complex phenomenon influenced by a large range of factors. Second, it is not known to what extent sugar consumption as a whole has changed. Like our hypothetical researchers in the UK, they relied on extant data sources (a survey of household food and drink in urban areas) for their outcomes. It is possible that people would have reduced consumption anyway, that they are consuming more homemade *agua fresca* (sweetened fruit infusions popular in Mexico), or that they have taken to eating sugary snacks instead.

**Rhetoric and ritual**

Policy, explanation and research are all linked, and the three relationships play a role. First, there is a mutually beneficial relationship between researchers and policy-makers that can have an adverse impact on research findings. Policy-makers value scientific research findings as a basis for policy because the public trust scientists more than politicians to tell them the 'truth' about whether a given policy is a good idea or not (Ipsos MORI, 2016). Researchers in turn welcome the opportunities from government-funded research. But rhetoric, particularly in politics, is weakened by any indication of uncertainty, and the caveats that form part of many scientific publications are omitted wholesale at the policy stage.

Secondly, researchers do not always take control of explanatory inferences, which leaves this task open to others after the fact. In the absence of local, recent, relevant evidence, policy-makers can place the burden of proof either on the proposed change (because its effects are not well known) or the status quo (because it is generally agreed to be imperfect); both are logically sound. This makes the relationship between research and explanation a crucial one, and as explained above, simple statistical inference falls short of explanation in complex systems.

The third relationship, between explanation and policy, is contested as both the researcher and the policymaker can offer explanations, and these can reflect biases and interests. Seeking a better explanation by more detailed study is time-consuming and expensive, making it unappealing to the funders as well.

Policy decisions affecting public health are often binary (in our example, the tax is either charged or not). However, the mechanisms by which the tax affects behaviour are far from binary, and research needs to provide understanding of these mechanisms, not just an average. Researchers will have to explain that it is unrealistic in a complex system to expect an effect invariant over time, space, individual circumstances or other contexts, and so understanding a variety of mechanisms by which change comes about is more important than attributing a single number to it. With this insight to hand, the policy maker can still reduce the statistics to an average effect, but will now be aware of whether this might in fact have an unexpected disadvantage in some settings.

**Our tentative solutions**

We contend that explanations of why particular transient patterns have occurred are vitally important. A similar set of circumstances may recur in the future, with the system responding in a similar way, and this is impossible to identify from reduced-scope, simplified or strong-assumption analyses. If the role of the analysis is in some way to predict the future status of the system, then both detailed quantitative and qualitative descriptions will be needed. For this reason, we regard mixed-methods research as essential for complex systems. Why make strong assumptions about

the relationships among the agents when you could simply ask them? Conversely, why simply record what they tell you (and trust them) when you could also make objective measurements of the effects? Only by combining the two can statistical and explanatory inference be brought together in studying complex systems. The explanatory inference may help to identify recurring circumstances in the future, or it may not, but it has a better chance of useful predictive ability than the abstracted statistical inference alone.

Attempting to systematise the role of explanation alongside statistical and qualitative inferences is a considerable challenge. If practical approaches can limit the scope for cryptic multiplicity, they can then be extended to the opaque one-to-many mapping of statistical inference onto explanations. Leamer made suggestions including 'fragility analysis', which is related to Platt's 'strong inference': to set out a pre-specified plan to analyse your data on the basis of your explanations, and then also those of people with opposing prior beliefs. If the conclusions are unchanged by this, they are not fragile (Leamer, 1983; Platt, 1964). Leamer, Gigerenzer and Manski, each writing at the interface between research and policy, have all urged researchers to make more thorough pre-specified plans, including how one would respond to particular patterns in the data by cleaning, checking, introducing more detailed models or investigating subgroups (Gigerenzer, 2002; Leamer, 1983; Manski, 2013).

This would help but would be burdensome and difficult. Nevertheless, we are optimistic, because clinical trials now routinely use registration, oversight and pre-specification. The Medical Research Council in the United Kingdom has published advice on evaluating complex interventions, which is relevant to the problems we are considering here, although we believe that this guidance needs to go further, requiring detailed pre-specification and be backed by monitoring of the details of analysis (Moore et al., 2015). As Platt pointed out, the most effective scientific discovery seems to proceed by a series of focussed deductive experiments in an iterative, inductive process (Platt, 1964). The pre-specification of explanations is not possible over the longer inductive process but applies to the deduction.

Perhaps all research in complex systems requires a data monitoring committee, such as clinical trials have, which considers not only the conduct of - and possibility of halting - the data collection, but also the conduct of - and possibility of halting - the analysis. How then can this be extended to explanation? Lipton characterised a good explanation as both likely (matches all relevant data) and lovely (provides new insight into why something happened, especially if phenomena outside the study are also explained) (Lipton, 2004). One way of avoiding cryptic multiplicity might be for researchers to pre-specify what they might regard as likely and lovely explanations within a given range of results. Lipton goes further and suggests that Bayesian analysis could explicitly incorporate prior probability distributions representing likelihood and loveliness, which is intriguing but requires methodological development (Lipton, 2004).

Education and training will be vital in improving the understanding of inference. The American Statistical Association's Guidelines on Assessment and Instruction in Statistics Education provide a framework that we believe would help many researchers to achieve a deep understanding of what they do beyond a 'cookbook' of statistical recipes that deliver incredible certitude (Aliaga et al., 2010). This is especially important for those taking courses such as Masters of Public Health, where study design and statistics is necessarily a relatively small part of the course and the cookbook is a tempting way to cover standard topics quickly.

Recent reviews of research using complexity theory have criticised the lack of mathematical models. We are wary of deterministic mathematical models reintroducing old statistical problems, but microsimulation, also called agent-based modelling, a technique from economics, offers a very flexible framework that could help with understanding and explaining patterns in complex systems (Greenwood-Lee et al., 2016; Thompson et al., 2016).

Realist evaluation is a valuable framework on which to build more effective research in complex systems (Pawson, 2013), particularly in the sense that learning about the system is an ongoing effort and may require adaptive interventions with ongoing evaluation (Manski, 2013). We do not believe that this precludes positivism or need usher in self-perpetuating research high on buzzwords and low on content, as some have critiqued (Cilliers, 2005; Thompson et al., 2016). Normalisation process theory is a similar recent effort, more focused on health systems (Murray et al., 2010).

**The near future, revisited**

> In our imaginary scenario, two years have passed since the sugar tax was rolled out nationwide. A new assessment by different researchers has criticised the earlier study: some locations show bigger reductions than others, there is some evidence of effect modification by socio-economic class, along with a shift to other sources of high-sugar drinks, and the original results may have been biased by creating unintended incentives on the school nurses and supermarket managers. Because these pieces of research were conducted with different designs and populations, nobody can judge between them, and because the tax is a *fait accompli*, there is no appetite to fund further research.
>
> It could have been different. Imagine our hypothetical researchers beginning by seeking explanation and not just a statistical contrast. Recognising the sugar tax as a complex intervention, their project is planned using recent concepts of evaluation (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012; Craig et al., 2008; Green et al., 2015; Murray et al., 2010).
>
> A mixed methods project asks shopkeepers, wholesalers, schools, health professionals and the public about how they regard the tax as functioning.

Explanations for local contextual effects are sought, and the researchers control the message by refusing to provide a statistical interim report but rather a qualitative one, with detailed mixed methods reports to follow. The politicians like the idea of individuals' stories followed by statistics.

The researchers warn that there are likely to be changes over time, as consumers and businesses adjust to the tax, so a snapshot could result in embarrassment later. They also list a number of ways in which a single effect might not apply across the country, or in other contexts like work or school canteens, and online bulk sales. They anticipate the food industry adapting with products that fall outside the definitions in the new law. They plan a priori subgroup analyses with linked interviews to look for sugary drinks becoming paradoxically popular, for example as a macho status symbol among teenage boys.

The researchers negotiate to obtain excise data to augment that from retailers and households. They expect response bias to the latter and include this in a Bayesian analysis. They also argue for the tax to be phased in at different rates, so that a dose-response relationship can be studied. Finally, they demand that interpretation is not done behind closed doors. They could plan for these problems because they were suggested from an initial round-table discussion to explore possible explanations.

This ending to our hypothetical scenario is deliberately optimistic but reflects how some of our suggestions could be implemented.

**Conclusion**

We have reviewed some diverse opinions on the shortcomings of research under two difficulties: examining complex systems and making an explanatory inference rather than just a statistical one. We contend that this double difficulty has contributed to the so-called crisis of replication critiqued first in psychology and social science but now more widely. Although much has been written about these individual difficulties, and some solutions proposed, here we bring together inference to the best explanation and the gap between statistical and explanatory inferences, to show how this manifests in cryptic multiplicity and the crisis of replication, particularly in complex, adaptive systems.

It is possible to carry out research in a complex, adaptive system in a principled way, influencing policy appropriately and providing reliable predictions. Much of the literature our argument is based on is already known to public health experts but does not seem to be regularly used. Moreover, there is a danger that explanatory inference, particularly if it is associated with the qualitative part of a mixed methods study, continues to perform 'a largely auxiliary role in pursuit of

the technocratic aim of accumulating knowledge of “what works”’ (Howe, 2004, p.53, as cited in Denzin and Lincoln, 2005, p.7). Addressing this issue is far harder work than just requiring a theory of change to accompany a headline statistical contrast, when it is the latter that attracts the attention of policymakers and promotes the uncritical progression from ‘pilot’ to ‘roll-out’ of a public health initiative. Much methodological work is yet needed before the integration of statistical and explanatory inference reaches maturity.

**Funding**

This work was not supported by any funding.

The authors have no financial, business or other interests to declare.

**References**

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., … Witmer, J. (2010) *Guidelines on Assessment and Instruction in Statistics Education: College Report.* American Statistical Association. Retrieved from: http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf

Baker, M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature,* 533, 452-454.

BBC (2015). *Today*, Radio 4, 11 Dec 2015. Transcribed from online recording by RLG.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., Moore, L. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75, 2299-2306.

Byrne, D. (1998). *Complexity and the Social Sciences*. Abingdon: Routledge.

Campbell, N., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., … Kinmonth, A.L. (2007) Designing and evaluating complex interventions to improve health care. *British Medical Journal*, 334, 455-9.

Cilliers, P. (2005) Complexity, Deconstruction and Relativism. *Theory, Culture & Society*. 22, 255-267.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*, 337, 979-983.

Colchero, M., Popkin, B. Rivera, J. & Ng, S. (2016) Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: observational study. *British Medical Journal,* 352:h6704. doi: 10.1136/bmj.h6704

Denzin, N.K., & Lincoln, Y.S. (2005). *The SAGE Handbook of Qualitative Research (4th edition).* London: Sage.

Donaldson, S. (2007). *Program theory-driven evaluation science: Strategies and applications*. Mahwah: Erlbaum.

Gelman, A. & Loken, E. (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. Unpublished manuscript. Retrieved from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gigerenzer, G. (2002) *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Green, J., Roberts, H., Petticrew, M., Steinbach, R., Goodman, A., Jones, A., Edwards, P. (2015) Integrating quasi-experimental and inductive designs in evaluation: A case study of the impact of free bus travel on public health. *Evaluation*, 21, 391-406.

Greenwood-Lee, J., Hawe, P., Nettel-Aguirre, A., Shiell, A. & Marshall, D.A. (2016) Complex intervention modelling should capture the dynamics of adaptation. *BMC Medical Research Methodology,* 16 - 51.

Harvey, D. L. & Reed, M. (1996). 'Social science as the study of complex systems'. In Kiel, D. L. & Elliot, E. (eds.), *Chaos Theory in the Social Sciences: Foundations and Applications.* Ann Arbor: The University of Michigan Press.

Hood, R. (2014). Complexity and integrated working in chioldren's services. *British Journal of Social Work*, 44, 27-43.

Ipsos MORI (2016). *The Ipsos MORI Veracity Index 2015: Trust In Professions*. Retrieved from https://www.ipsos-mori.com/researchpublications/researcharchive/3685/Politicians-are-still-trusted-less-than-estate-agents-journalists-and-bankers.aspx

Leamer, E.E. (1983) Let's take the con out of econometrics. *American Economic Review,* 73, 31-43.

Lewin, S., Glenton, C., Oxman, A.D. (2009) Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. *British Medical Journal*. 339:b3496. doi: 10.1136/bmj.b3496

Lipton, P. (2004) *Inference to the Best Explanation (2nd edition).* Abingdon: Routledge

Manski, C. (2013) *Public Policy in an Uncertain World: Analysis and Decisions*. Boston: Harvard University Press.

Mayo, D.G. & Cox, D.R. (2006) Frequentist Statistics as a Theory of Inductive Inference. *Institute of Mathematical Statistics Lecture Notes - Monograph Series.,*49, 77-97

Moore, G.F., Audrey, S., Barker, S., Bond, L., Bonell, C., Hardeman, W., Baird, J. (2015) Process evaluation of complex interventions: Medical Research Council guidance. *British Medical Journal.* 350:h1258. . doi: 10.1136/bmj.h1258

Murray, E., Treweek, S., Pope, C., MacFarlane, A., Ballini, L., Dowrick, C., May, C. (2010) Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. *BMC Medicine*, 8:63. doi: 10.1186/1741-7015-8-63

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. London: Sage.

Platt, J. (1964) Strong Inference. *Science*, 146, 347-353.

Sayer, A. (1992). *Realism and Social Science*. London: Sage.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science.* 22, 1359–1366.

Stacey, R. D. (2007). *Strategic management and organisational dynamics: The challenge of complexity to ways of thinking about organisations (Fifth edition).* Harlow: Pearson.

Thompson, D.S., Fazio, X., Kustra, E., Patrick, L. & Stanley, D. (2016) Scoping review of complexity theory in health services research. *BMC Health Services Research*, 16, 87.

Wight, D., Wimbush, E., Jepson, R., & Doi, L. (2015) Six steps in quality intervention development (6SQuID). *Journal of Epidemiology and Community Health*. Advance online publication. doi: 10.1136/jech-2015-205952