The role of value judgment in the production of evidence for policy: lessons from development interventions in health and nutrition.

MUNSLOW, TAMLYN

How to cite:

MUNSLOW, TAMLYN (2024) *The role of value judgment in the production of evidence for policy: lessons from development interventions in health and nutrition.*, Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/15798/

The role of value judgment in the production of evidence for policy: lessons from development interventions in health and nutrition

Tamlyn Munslow

Abstract

This thesis concerns the production and use of evidence, especially for policy improvement. It is my main thesis that science cannot avoid making value judgments when it comes to issues of evidence for policy use. But, more importantly, that value judgments even inform the drawing of conclusions on the basis of evidence. Whether or not the production of evidence can and should be done in a value-free way is a highly controversial claim involving many disputes. What matters is that the total body of evidence for evaluating a particular claim considers evidence for and *against* a claim from different relevant value perspectives. This requires a concerted effort to investigate all the best available evidence about a preferred course of action. This is an interdisciplinary topic, and I address it in an interdisciplinary way, with arguments from philosophy of social science, social policy, and political science, including the subdiscipline policy studies.

The role of value judgment in the production of evidence for policy: lessons from development interventions in health and nutrition

Tamlyn Munslow

A thesis submitted to the Department of Philosophy at Durham University for the degree of Doctor of Philosophy, June 2021

Table of Contents

Table of Contents

Declaration

I certify that the thesis I have presented for examination for the PhD degree of Philosophy at Durham University is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

Acknowledgements

I would like to thank my primary supervisor, Professor Nancy Cartwright, for her continuous support and wonderful guidance throughout my PhD. I have learnt such a huge amount from Nancy and feel incredibly privileged to have her as my supervisor.

I would like to thank my secondary supervisor, Dr Wendy Parker, for her fantastic support and guidance throughout my PhD.

I would also like to thank my husband, Sean Connell, for always believing in my PhD.

Further, I would like to thank, in no particular order, Adrian Harris, Alison Wylie, Eileen Munro, Louise Caffrey, Haydon Munslow, Richard Williams, William Peden, Donal Khosrowi, Katherine Furman, Erin Nash, Andrew Fletcher, Julian Reiss, Sarah Wieten, Dave Byrne, Doug Scott, Jeremy Hardie, Robin Hendry, Karen Munslow, Eleonora Montuschi, Peter Vickers, Nicola Craigs, Rob Van den Berg, Elliot Stern, Uzma Malik, and Omar El Mawas for their thought-provoking contributions to the topic of values in science. I would also like to thank members of the Centre for Humanities Engaging Social Science (CHESS) and Knowledge for Use (K4U) research groups at Durham University for imparting such wisdom.

Finally, I would like to express my gratitude for the extensive financial support that I have received in the form of a European Research Council (ERC) Doctoral Studentship as part of Professor Nancy Cartwright's Horizon 2020 funded Knowledge for Use (K4U) project. Knowledge for Use (K4U) is an innovative, interdisciplinary, and multi-institutional ERC-funded research project that aims to show how to put scientific research and common knowledge together to build more decent societies.

For my daughters

Chapter 1: Introduction

My thesis concerns the production and use of evidence, especially for policy improvement. Whether or not the production of evidence can and should be done in a value-free way is a highly controversial claim involving many disputes. I am concerned with whether and in what ways contextual values inform the very core of scientific investigation, in particular by playing a role in the choices about what to research, what methods to use, and what conclusions to draw, where I shall be especially concerned with the collection, evaluation and use of evidence.

First, there are some important distinctions to make. There are important terms used throughout this chapter, including science, research and evidence. Care has been taken to make sure the use of this language is consistent.

- *Science* refers to the systematic study of the structure and behaviours of the physical and natural world primarily through the means of observation and experiment.
- *Research* refers to the systematic investigation of materials and sources in order to establish facts and reach new conclusions.
- *Evidence* refers to facts or information indicating whether a belief or proposition is true or valid.
- *Policy intervention to achieve designated outcomes* refers to actions and activities undertaken by a respected body (often national or international authorities) to accomplish those outcomes. Policy interventions address a variety of measures including technologies, practices and behaviour. This includes, for example, the provision of nutrition monitoring in countries where undernutrition remains a key policy concern.

It is my main goal to convincingly argue that values play an important and ineliminable role in various ways and at various points when science is used to inform social and health policy intervention, including the prediction of intervention outcomes, the formulation of the intervention, its implementation and its evaluation. The value ladenness of science has been a central topic in the philosophical literature for a while now. Some proponents of the value ladenness of science take value ladenness of some sort or another to be endemic across the sciences. I do not here take any stand on values in science in general. Rather, I will call on this literature where it seems especially relevant to my concerns, bringing some of the arguments from this literature to bear on the value ladenness of science as used to inform social and health policy.

Some proponents who use arguments from underdetermination and inductive risk share the premise that we should only consider values where the evidence runs out or leaves uncertainty. Their primary focus is on evidence being used to reach scientific conclusions while avoiding any situation where conclusions may be reached on the basis of 'wishful thinking'.

By using science to inform policy, the idea is that we can make decisions informed by scientific evidence. In order to have a greater impact on health policy, for example, we can base policies on evidence that are supported by scientific methods. What my thesis does is make explicit some places where value laden assumptions can play a role in the collection and processing of evidence for a particular claim, as well as make clear what role those values are playing, and what the implications are. I will argue that values have a pervasive influence, and that this needs to be acknowledged (see also Munro, 2013, Montuschi, 2013, Cartwright, 2013, etc). As a way of contributing to these debates, I look at the production and use of evidence for policy as a three-step process:

1. Finding out what facts exist
2. Identifying reasons to see those facts as evidence
3. Understanding how evidence is woven together to confirm or disconfirm a claim.

A particular type of evidence that I discuss is evidence for the evaluation of a claim that a policy or intervention worked or will work as intended. I argue that what matters is that the total body of evidence for evaluating a particular policy claim should consider evidence for and *against* a claim from different relevant value perspectives. This requires a concerted effort to investigate all the best available evidence about a preferred course of action. This is an important topic if we are to better understand the specific role of values in the production and processing of evidence for policy improvement. This may seem totally uncontroversial when expressed explicitly. As I note in chapter five, it is a central part of Bayesian confirmation and the related methods for case study inference. Nevertheless, it is very often ignored in practice in the evaluation of policy claims. I shall show how easy it is to fall into this practice because the requirement to gather and compare the evidence both for a hypothesis and for plausible contrary alternatives is not explicitly mentioned in many evaluation methodologies—including one I myself helped elaborate, which I discuss in chapter five.

The evaluation of policy claims is a specific part of the policymaking process. This is a key area of focus in this dissertation. Policy evaluation typically applies a range of research methods to systematically examine the implementation or impact of a policy intervention. Evaluation is the main activity through which we develop an understanding of the merit, worth, and value of such a policy in terms of improving the social and economic conditions of the different stakeholders. It is a widely held view that evaluation can (under the right conditions) support the policymaking process by providing useful recommendations about the implementation or impact of a policy.

It is my main thesis that science cannot avoid making value judgments when it comes to issues of evidence for policy use. But, more importantly, that value judgments even inform the drawing of conclusions on the basis of evidence. So, when we say that there is evidence of an improvement in health or wellbeing, for example, we must understand that we commit to certain value-laden assumptions in offering those statements. Facts don't come with a label on them saying, 'I am evidence for a particular claim'. Rather we need to have reasons to use various facts as evidence in support of (or against) a claim. What is and what is not evidence, and for what claim, is underdetermined by the facts, and so we draw on background assumptions to turn facts into evidence for assessing claims. Background assumptions will sometimes reflect values, including moral, political, cultural, and other points of view (Longino, 1979; Atkinson, 2001; Anderson, 2002; Di Bucchianico, 2009; Crasnow, 2014).

I aim to understand what we mean when we say that values inform science. According to the value-free ideal, the internal workings of science, including the evaluation of evidence, should be kept free from the influence of non-epistemic values (or contextual values--Longino) as much as possible. Issues of how to understand the role of values in science and the responsibilities of scientists are often excluded from the study of knowledge (Douglas, 2009: 44). Most of this exclusion has been on the basis that science is (or should be) a value-free enterprise and that scientists should not consider the broader implications of their work when conducting scientific research (Douglas, 2009: 44). Such a basic view of science is inadequate for understanding the production of evidence for policy (Douglas, 2009).

By 'contextual values' I mean to include a broad range of influences which include, but are not limited to, the social and cultural environment where science is done, which includes the norms, preferences, beliefs, or interests about what is acceptable or unacceptable practice, and which vary from community to community (Longino, 1990). These may, and often do, include what we would normally judge morally relevant norms, preferences, etc. – like the preference to serve one social group (perhaps oneself and one's family) over others.

Since contextual values can bias our conclusions, several contextual empiricist and feminist philosophers of science urge that we must be explicit about *where* they enter (Longino, 1990; Anderson, 2002; Intemann, 2005). Many feminist philosophers commit to the view that science neither is, nor ought to be, neutral among social and political values. They say that the necessity of background assumptions mean values are unavoidably entering our inquiries and so we need to set up our practice to avoid some biasing assumption from dominating. By taking stock of other value perspectives, for example, we can mitigate against biasing assumptions (Longino, 1990). Some feminists offer a further role for the inclusion of different value perspectives in the production of evidence. This extends beyond the woman's perspective and includes representation from other underrepresented individuals, like non-western opinions or other often marginalised perspectives (Intemann, 2010). I bring in insights from the field of international development to bring light to a more persistent problem discussed by some feminist philosophers of science: the problem of actively reproducing the ideas that damage vulnerable communities and reinforce current injustices.

I suppose that values come from the social and political contexts in which research and evaluation takes place, and it is inevitable that contextual values will influence the processing of evidence, given the problem of underdetermination and the fact that evidence processing must always take place within some conceptual scheme or another which often can reflect contextual values. Hence, values often inform our efforts to assess scientific claims and concepts -- here I will be concerned with scientific claims and concepts that are used to inform policy -- whether we strive to make these assumptions explicit or not. The production of evidence is located in a political context, and ignoring this, or failing to understand what the implications of this are, risks allowing presumptions or vested interests to drive evidence towards a predetermined conclusion, and further risks legitimising one set of ideas over another. We need to understand how facts are produced, what facts hide or emphasise, and to better understand how we might consider alternatives.

This is an interdisciplinary topic, and I address it in an interdisciplinary way, with arguments from philosophy of social science, social policy, and political science, including the subdiscipline policy studies, among others. I draw on diverse examples from within the international development literature to inform my research. International development is a particular field of study where we may expect to find diverse contextual values in the beliefs, preferences, and interests of different players. The field combines different disciplines and various methodologies, including qualitative, quantitative, and mixed methods approaches. Different disciplines often disagree about what is acceptable or unacceptable methodological practice and differ in their beliefs about the extent to which, and ways in which, diverse communities should be involved in the production of evidence for use in policy. Some of the literature about values in the international development field seems to lack clarity as the term values is applied in many ways. I want to consider in some detail different ways that we might want to think about values so that we can better identify where values are influencing, and what the implications are.

My examples are informed by previous work produced by the Institute of Development Studies. These examples are open access. The Institute of Development Studies programme on Strengthening Evidence-based Policy works across different policy themes. Each theme works with partner institutions to co-construct policy-relevant knowledge and engage in policy-influencing processes. In doing so, the research actively seeks to include representation from underrepresented individuals, like non-western opinions and other marginalised perspectives, working in partnership with stakeholders closest to the phenomena under investigation. An Open Access publication distributed under the terms of the Creative Commons Attribution License permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are clearly credited.

My three main examples include: the World Health Organisation (WHO) response to the 2013-2014 Ebola outbreak in West Africa; the monitoring of political commitment to address hunger and undernutrition (te Lintelo et al. 2020); and the surveillance of infant undernutrition (Barnett et al. 2020). These examples focus broadly on disease and nutritional surveillance, which involves the setting up of systems to monitor progress in health and nutrition to make decisions that lead to improvements in target populations. My examples pay attention to the priorities and assumptions inherent in a policy intervention, but also of the clients of those interventions, including the communities and extended families of those affected.

These are important and timely topics. The surveillance of infant undernutrition encompasses a variety of programmes that utilise smartphone technology with the aim of supporting governments to transform and modernise community health and nutrition services. What’s interesting is that despite the predicted success of real-time data and the technologies that promote it, the use of technology in nutrition services is still relatively new, and it is inconclusive whether the use of technology leads to improvements. What my examples show is that the potential of new technology and the data produced is huge, but it is unwise to take excessive pride in impressive technology since assumptions about its benefits to users are often not met. These are ideal cases to explore the ways in which contextual values inform the very core of scientific investigation, in particular choices about what to research and what conclusions to draw.

“The inevitability thesis”

My thesis contributes a variety of cases that clearly illustrate the inevitability of values in the production and use of evidence for social policy in particular. My dissertation suggests a strong “inevitably thesis” to the effect that values inevitably influence the choice and use of scientific evidence, methods and concepts when they are employed to inform policy. At the onset, I need to clarify the scope of this inevitability thesis, restricting it to social policy (and excluding e.g., string theory).

I will first illustrate this point with an example. Consider the scientific claim: weight loss (or insufficient gain in weight) in infancy leads to ill health throughout the life course. Weight is a chief indicator of deteriorating health in children, especially those under 2. Weight loss is therefore something society justifiably wants to avoid. For example, the first 1000 days are critical, and we need to monitor weight, as a decline in weight can increase the likelihood of disease in later life, which we obviously do not want. Weight as a proxy for poor nutrition is incontestable. However, weight is still not a value neutral proxy as it reflects wider assumptions about what is feasible to monitor, what is important, what is significant, and this will always depend on a wider social purpose. Hence, the choice to measure weight (and not something else) is value laden.

Of course, one example does not by itself establish a general claim like mine. But the features of this example that matter to my claim are general—the fact that social policy is meant to address things we care about and the details of what are considered to constitute ‘what we care about’ for any given case will affect the claims that are relevant.

Consider another scientific claim: breastmilk promotes the healthy growth of a child and ought to be given exclusively. What this tells us is that breastmilk is the best type of nutrition for the child. Hence, community health clinics might seek to produce data on the number of mothers receiving breastfeeding counselling and other services when a cohort of children’s weights seems to be declining. Producing data on the number of mothers attending counselling services is a useful indicator of infant health as it is demonstrative of a likely improvement in feeding practices since we want infants to have the best source of nutrition and not simply to put on weight. However, producing data on the number of mothers in counselling services is value laden because it is data derived for a social purpose (that we *want* infants to have the best possible start in life). Hence, sometimes there are matters in social policy that are obviously incontestable; like the choice to monitor weight and the choice to monitor mother's attendance at health clinics. However, we produce data on these indicators not because they offer a value neutral means to monitor undernutrition in children; but because society doesn't want children's weight to drop and cause unavoidable and likely harm to the child. This is not value neutral.

Consider again the decision to monitor the number of mothers in counselling services. This produces data on the relative number of mothers attending the sessions but offers no means of understanding the satisfaction of those mothers, or the wider influences shaping the mother's ability to nourish her child. Choosing to focus on one type of data affects the collection of data on alternative proxies of infant health (e.g., mothers' buy-in to the services offered by the clinics). We need to understand how values are influencing the production of knowledge to understand the ways in which some interests shape nutrition policy over others.

These are relatively straightforward examples. In my thesis, we will delve into more complex examples from wider areas of social policy. However, the point I wish to make is as follows: it is inevitable that values will influence the production and use of knowledge to inform social policy. That said, there are examples where this is relatively unproblematic (like the decision to monitor weight as a proxy of growth), and there are examples where this is more problematic.

What is needed is clarification of exactly *how* values are involved in the production and use of evidence for policy. This is important since we inevitably end up making judgments about what is important, what ought to be studied, what result is of practical significance, and to whom. I make a contribution to these topics in the following way:

Chapter two is primarily a review explaining and clarifying a body of current literature that is relevant to my concerns about the role of values in policy prediction and evaluation. It lays out an argument, predominantly by Heather Douglas, that science can be value-laden yet retain its integrity as a source of reliable knowledge. Douglas proposes that, rather than limit the *kinds* of value that can legitimately play any role at different stages of research, we should limit the *role* that any value can legitimately play at different stages of research. For example, one way that it is appropriate to use values at any stage in evaluating the acceptability of a hypothesis is to assess the *inductive risk* of one procedure over another and use that assessment to decide what to choose. Douglas offers a role for values in judging whether there is sufficient evidence to warrant the acceptance of a hypothesis, given the unintended consequences of error (Douglas, 2000). But Douglas does not want values to enter in deciding whether evidence *is* evidence for a hypothesis, nor in deciding how strongly the evidence supports the hypothesis. While others argue that values of some kind or another are inevitable in these two processes as well. This is one of the places I hope this thesis contributes, by first, clarifying how values are inevitable in these two processes when it comes to policy prediction and evaluation and what difference that makes and second, providing detailed illustrations in real-life health policy cases of serious recent concern (Munro, 1999, 2005; Intemann and de Merlo-Martin, 2010).

Chapter three lays out an argument, predominantly by Helen Longino, that makes the role of *background assumptions* explicit in science. This is the argument that background assumptions enter in deciding what is evidence for a hypothesis and then background assumptions are needed in deciding whether to accept the hypothesis given the evidence (Longino, 1990). I offer examples from feminist empiricist philosophers who claim a role for contextual values in health science (e.g., Longino, 1990, 2002; Anderson, 2002, 2004; Crasnow, 2014). I also consider a new example of the 2014 West African Ebola outbreak (e.g., Wilkinson and Leach, 2015; Fairhead, 2016). What I intend to illustrate with my examples is how a commitment to different value perspectives leads us to focus on different evidence. This is with the aim of understanding how and where values may enter science and what the implications might be. I defend a particular role for contextual values in the collection and processing of evidence: in the selection and collection of facts, the rendering of facts as evidence and the weaving of these together to confirm or disconfirm a hypothesis. This chapter uses a mix of considerations from both the philosophy literature and well-known causal-process tracing methodology.

Chapter four considers the use of value-laden assumptions in measurement. By measurement I am referring to the conceptualisation of concepts that are of interest to us in social science (Bradburn, Cartwright and Fuller, 2016). I lay out a new example of the construction and use of the Hunger and Nutrition Commitment Index by researchers at the Institute of Development Studies. The Hunger and Nutrition Commitment Index is a particularly interesting example as researchers recognise that measures of political commitment may conceal underlying assumptions about social change and so seek to remain transparent in their methodological choices and their consequent impact on the rankings (te Lintelo et al. 2020). One essential driver of political commitment is the strength of nutrition actors. For example, at the international level, networks have proliferated in recent years as distinct organisational forms to address several high-burden health issues. One key example is the Scaling Up Nutrition (SUN) Movement, which is the world's largest initiative for building commitment and scaling-up multisectoral implementation efforts. The researchers acknowledge the centrality of this network in mobilising political commitment and utilise a co-production methodology that is designed to include members as partners in the production of evidence on political commitment. They propose that equitable partnerships between producers and users will enhance uptake among those users, in turn increasing the likelihood that that evidence informs policy. Practically, this may improve the effectiveness of generating and sustaining political commitment among diverse actors. Normatively, we can ask what biases – potentially in favour of larger, well connected, and professionalised organisations – this introduces into the policy process.

Chapter five concerns the production of evidence on how programmes are supposed to produce their targeted effects, as is now widely recommended in the international development community. Programme theories refer to a structured set of assumptions about how an intervention is expected to work and how it is expected to influence processes of change. To be useful for the purpose of informing policy programme theories must be presented in such a way as to allow policy specialists to understand the facts of the case.

This serves the purpose of producing useable evidence for policy specialists that are derived from the facts of the case and is widely recognised as a crucial part of an evaluation process.

This chapter describes original research that I completed in collaboration with several others that has already been published in the form of Working Papers for the Centre for Excellence in Development, Impact and Learning (see Cartwright et al. 2020). This is an interdisciplinary centre that was launched by the then Department for International Development (DFID). In a recent paper, 'Making predictions of programme success more reliable', Cartwright et al (2020), I, along with the other authors, set out the role of *causal-process-tracing theories* for evidence gathering and assessment. Our paper is intended to serve as a useful a template for categorising evidence for policy. We offer a template that presents the step-by-step process by which an intervention is to produce its intended effects, including specific information about the "principles of change" for each step and the implications of these (Cartwright et al. 2020: 7). During this project, I noticed a major missing ingredient in the kinds of evidence represented that everyone would agree should definitely be part of the evidence base for policy prediction and evaluation. That is the requirement that the total body of evidence for evaluating a particular claim should include not just evidence that directly confirms or disconfirms this claim, but also evidence that rules out alternative theories of what will happen. I aim to make this requirement explicit in my chapter.

What possible evidence gets noticed and included will be heavily influenced by the cultural values of those proposing the causal-process-tracing theories and of the setting in which they work. What I intend to show is the need to engage with different value perspectives to identify priority issues (e.g., ethical, social, political) to support the refinement of causal-process theories. I tackle head on the collection of evidence to rule out alternatives and propose that this needs to be added as a separate item besides the kinds of evidence that appear in our joint Working Paper template.

Chapter six concludes by offering implications for a wider project into how to deal with values in the production of evidence for policy. One issue that my thesis demonstrates is the problem of committing to a particular value perspective in our attempts to understand a particular phenomenon. A single value perspective restricts how we understand something like health or undernutrition. The problem is not that the account on offer is untrue, but that it is incomplete. Developing ways to avoid single value perspectives when conducting research and evaluation for policy is a key area for further study. Of relevance are the distinctive priorities, motivations, and concerns of disparate stakeholders in research. We need to understand the different values and understandings brought to the study of health and undernutrition by major stakeholders: patients, their caregivers, clinical staff, civil society groups, umbrella organisations, businesspeople, policy makers and people from the media. In addition, we need to ask what biases major stakeholders introduce into the policy process. Much work has already been conducted on this topic, yet it remains an important and timely interdisciplinary topic with arguments from philosophy of social science and practical empirical based research.

Of course, there are many questions relevant to the topic of values in science that this thesis does not ask. I point out that value judgments are inevitable – choices need to be made and they will have value implications regardless of whether anyone deliberates about which should dominate. I do not ask whether a particular programme goal is the *right* one, and for whom? I do not ask who 'should' make the value judgment. These are important topics, but it was important for the aims of this thesis to focus on the following areas: what values inform science, where values play a role, and what the implications are.

References

Atkinson, A. B. (2001) The Strange Disappearance of Welfare Economics, *Kyklos*, 54, (2/3), 193-206.

Anderson, E. (2002) Situated Knowledge and the Interplay of Value Judgments and Evidence in Scientific Inquiry, *In the Scope of Logic, Methodology and Philosophy of Science*, 497-517.

Barnett, I., Sulistyo, S. S., Befani, B., KariSari, K., Sharmin, S. and Dewi, D. 2016. *Mixed-Method Impact Evaluation of a Mobile Phone Application for Nutrition Monitoring in Indonesia*. Institute of Development Studies.

Bradburn, Cartwright, N. and Fuller (2016) A theory of measurement, Working Paper, Centre for Humanities Engaging Social Science and Society (CHESS), Durham.

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Williams, R.B. (2020), *Making predictions of programme success more reliable,* CEDIL Methods Working Paper. Oxford: CEDIL.

Crasnow, S. (2014) Feminist Standpoint Theory, Philosophy of Social Science: A New Introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford.

Di Bucchianico, Maria Elena (2009) *Modelling high temperature superconductivity: A philosophical inquiry in theory, experiment and dissent.* PhD thesis, London School of Economics and Political Science: United Kingdom.

Douglas, H. (2000) Inductive Risk and Values in Science, *Philosophy of Science*, 67, (4), 559-579.

Douglas, H. (2009) *Science, Policy, and the Value-Free Ideal*, University of Pittsburgh Press: Pittsburgh.

Fairhead, J. (2016) Understanding Social Resistance to the Ebola Response in the Forest Region of the Republic of Guinea: An Anthropological Perspective, African Studies, Review, 59, (3).

Intemann, K. and de Merlo-Martin, I. (2010) Social values and scientific evidence: the case of the HPV vaccines, *Biology, and philosophy*, 25, (2), 203-213.

Intemann, K. (2005) Feminism, underdetermination, and values in science, *Philosophy of Science*, 72: 1001-12.

Longino, H. (1990) *Science as Social Knowledge*. Princeton, NJ: Princeton University Press.

Longino, H. (1979) Evidence and hypothesis: An analysis of evidential relations, *Philosophy of Science,* 46, (1), 35-56.

Munro, E. (1999) Protecting children in an anxious society, *Health, Risk & Society*, 1:1, 117-127, DOI: 10.1080/13698579908407011.

Munro, E. (2005) Improving practice: child protection as a systems problem, *Children and Youth Services Review*, 27, (4), 375-391.

Munslow, T., Barnett, I. and Dewi, D. 2016. The Potential of M-health for Improved Data Use: Workshop Report, IDS Evidence Report 201. Brighton: Institute of Development Studies.

te Lintelo, D.J.H., Munslow, T., Pittore, K. and Lakshman, R. (2020) Process Tracing the Policy Impact of ‘Indicators’, European Journal of Development Research, 32, (4).

Wilkinson and Leach (2015) Briefing: Ebola Myths, Realities, and Structural Violence, African Affairs, 114, (454), 136-148.

Chapter 2: Inductive risk and values in science

2.1 Introduction

Values of some sort organise the methods of knowledge production. The question of where values come into science, and where they should and should not come in, has been given much consideration. One answer is they are revealed by our choice of assumptions, another the weight we give to different kinds of evidence and a third whether logic means strict rationality or reasonableness. Inductive risk arises when the evidence is insufficient to accept or reject a hypothesis. So, scientists face a risk of error whenever they accept or reject a hypothesis. According to the argument from inductive risk, as outlined by Heather Douglas, managing this risk properly requires considering non-epistemic values.

Philosophers distinguish between epistemic and non-epistemic values. Epistemic values are things like truth, approximate truth, empirical adequacy, internal coherence, and predictive accuracy. They are called "epistemic" because they are often taken as either conducive to, or as characteristic features of, scientific knowledge. Non-epistemic values are things like pragmatic reasons or values, social values, moral values, political values, religious values, aesthetic values. Non-epistemic values are indicative of normative or emotive commitments people hold. Different interests, like personal gain or concern for the advantage or disadvantage of particular groups, tend to get grouped under this use of "values" as well (e.g., one's clients or patients, minorities or "people like us"). It is the use of non-epistemic values that has been much harder to justify in scientific study.

In chapter three, I will go on to define what Helen Longino calls "contextual values" that can play in the process of scientific legitimation of hypotheses, which many have urged should be value-free (Longino, 1990: 216). However, this definition of values is not explicit in this chapter as this is not something that Heather Douglas commits to.

Heather Douglas and others (Longino, 1990) call the classical model of value-freedom "the externality model". On this model, values are external to the process of establishing scientific conclusions. There is an important distinction between activities that are “external” to the process of reaching a scientific conclusion and those activities that are “internal” or directly influence the scientific conclusion. Those activities that are external include the selection of a research question or scientific hypothesis, the choice of methodology, and the dissemination of scientific findings. It is less contentious that values of all kinds (epistemic and non-epistemic) do, and should, influence the external activities of science. Such as when we appeal to values when selecting a scientific hypothesis. What remains controversial is whether non-epistemic values have a role to play in testing a scientific hypothesis. Those activities that are internal, or directly influence a scientific conclusion, include the production of evidence to test a scientific hypothesis and the acceptance or rejection of a hypothesis based on the evidence. It has been much harder to defend the use of non-epistemic values in the justification of scientific findings (Douglas, 2000; Longino, 1990).

So, according to the external model of the value-free ideal, the justification of scientific findings should not be based on non-epistemic (e.g., moral or political) values. Heather Douglas and others have put forward different arguments against this version of the value-free ideal, defending the use of non-epistemic values in the justification of scientific findings. Douglas argues that scientists must employ moral judgments in managing inductive risks, and thereby finds a role for non-epistemic values in the very core of scientific investigation, at the point of accepting or rejecting a scientific hypothesis. Douglas does however go along with the value-free ideal in maintaining that scientists ought not to make value judgments in choosing what counts as evidence for or against a hypothesis and that values must never themselves be counted as evidence. That is what she calls a “direct role” for values internally in science. She maintains that direct roles undermine the integrity of science:

> … direct role for values, namely that values could be taken as reasons in themselves to accept or reject an empirical claim. It is this direct role for values that raises the double specter of self-deception among experts and the politicization of expertise. Constraining values to the "indirect" role allows one to maintain the integrity of expert reasoning, while allowing experts to behave responsibly towards their dependent direct role for values, namely that values could be taken as reasons in themselves to accept or reject an empirical (Douglas, 2000: 8).

But scientists can use value judgments “indirectly”, to consider whether the evidence is sufficient to warrant a particular claim given the consequences of wrongly accepting or wrongly rejecting the claim (Douglas, 2000).

> It is in weighing the importance of uncertainty that values, including social and ethical values, have a legitimate role to play in shaping our beliefs. This is what is meant here by an indirect role for values in expert reasoning (Douglas, 2008: 9).

2.1.1 Chapter outline

Douglas never gives up on the ultimate pursuit of truth, but she does offer a role for non-epistemic values to play alongside epistemic values. Douglas’s very influential work provides a very good framework for thinking about values in science, so I aim to review it in some detail in this chapter, laying it out clearly and accurately, although relatively briefly. I shall describe the categories that Douglas introduces of where we may find values entering in the practice of science and summarise her arguments about where they should and where they should not enter. I use examples from Douglas and others (e.g., Cartwright and Runhardt, 2014 and Sen, 1981) as illustrations throughout. But I add examples of my own concerning child welfare (Munro, 1999, 2005), climate change problems (Parker, 2014), and the development of vaccines (Intemann and de Melo-Martin, 2010) as part of my contribution to clarifying the issues.

I will use Douglas’ work to argue that values are unavoidable in the production and use of evidence to inform specific policies: e.g., child welfare policy, climate policy, and medical policy related to the spread of HPV. What is important is understanding that values can influence this process and deciding what to do about it. This statement suggests a strong “inevitability thesis” to the effect that values inevitably shape scientific evidence for use in policy. I restrict this claim to social policies, and do not make this claim for all areas of science.

2.2 The “externality model”

Within the externality model, values can be used to set the questions for investigation, to restrict the use of methods, and to make decisions about knowledge dissemination and use (Douglas, 2014). In this section, I will consider a role for values in those activities that are external to the process of reaching a scientific conclusion. Several cases inform this discussion including two that are widely cited and discussed in this regard. First, Milgram's Study of Obedience and the Tuskegee Study of Untreated Syphilis, as outlined by Douglas (2014). Second, the dissemination of findings relevant to kidney donations on BBC Panorama discussed by Amartya Sen (Sen, 1981).

2.2.1 Setting restrictions on scientific research

Consider Milgram's Study of Obedience carried out by psychologist Stanley Milgram at Yale University in 1961, as Douglas does (Douglas, 2014: 167). Milgram's experiment remains one of the most famous studies of obedience in psychology. The experiments began in July 1961, a year after the trial of Adolf Eichmann[1] in Jerusalem. Milgram devised the experiment to answer the question: Were Eichmann and his millions of accomplices in the Holocaust just following orders? Could we call them all accomplices?" (Milgram, 1974). Milgram was interested in whether the average individual would go against the authoritative scientist in the lab coat. Subjects were studied in a laboratory where they were told to administer increasingly intense electric shocks to another subject by a man in a white lab coat. Unfortunately, most of the subjects were compliant and administered what they took to be the most intense shocks (though in fact the subjects were not suffering – the participants were deceived in that). By contemporary standards, most people would agree that the risk to the psychological wellbeing of the subjects when they come to understand what they have done is too great to warrant the study.

Milgram, however, defended his experiment on the grounds that it offered new information on the nature of authority to obedience that would not otherwise have been possible (Douglas, 2014: 167). According to Douglas, Milgram claimed that the knowledge to be gained was worth the cost (Douglas, 2014: 167). The information was of benefit to society

---

[1] One of the major organisers of the Holocaust and the "Final Solution to the Jewish Question" in Nazi terminology.

since it offered a better understanding of the phenomenon of authority and how individuals' morality can become affected by it (Douglas, 2014: 167). Others have tried to defend the mistreatment of human subjects in certain cases, if a reasonable person would agree to the risks of the experiment (Reicher and Haslam, 2006 in Douglas, 2014; 168). Some argue that in unique circumstances, where the study design requires omission of the details that might alter participant response, it is possible that information about the study can be reasonably withheld from the participant until after their participation (Douglas, 2014). But, while in general it can be allowed that the gains from knowledge production can dominate over the participants' welfare, today it is commonly judged that this was not the case in the Milgram experiments (Douglas, 2014).

Douglas argues that whenever there is a judgment of whether to pursue a particular research question, a scientist—or some other responsible body—must weigh the value of knowledge to be gained against other risks (Douglas, 2014). This includes any risks to the research subjects because of the decision to pursue a potentially harmful line of inquiry. Douglas argues that Milgram's experiment approved an unethical procedure as it put the participants at the risk of psychological harm (Douglas, 2014: 167). Specifically, participants were purposefully deceived about the nature of the experiment. They were told that they were participating in an experiment about learning and pain, when in fact they were participating in an experiment on obedience to authority. Douglas argues that the protection of human subjects far outweighs any epistemic gain in Milgram's Study of Obedience (Douglas, 2014: 167-168).

For a second example, consider the infamous Tuskegee Study of Untreated Syphilis, as Douglas does. The study was conducted between 1932 and 1972 by the U.S. Public Health Service. The purpose of the study was to observe the natural progression of untreated syphilis in rural African American men in Alabama. Hundreds of low-income black men in Alabama were denied treatment for syphilis as part of a decades-long experiment to explore the progress of the disease. Some of the participants were also subject to painful treatments including lumbar punctures. The study is a renowned case because researchers knowingly failed to treat patients appropriately after the validation of penicillin in the 1940s (a treatment option for syphilis). There are serious ethical implications because information on possible treatment was consciously withheld from research participants, in some cases for up to 30 years. The study was pursued partly on the grounds that the experiments offered new information that would otherwise have been unattainable (Douglas, 2014: 167).

Cathy Gere puts this slightly differently. Doctors rationalised the non-consensual research on impoverished or undereducated human subjects on cost-benefit grounds: poor rural populations were otherwise out of reach of medical care; illiterate people could not be expected to understand risk. Gere argues that the doctors weighed up the risks and considered that the knowledge gained would be of advantage to the public, which meant that the doctors were free to pursue a potentially harmful method regardless of the risk to the human subjects (Gere, 2017).

Consider other examples, such as participants involved in Nazi research on hypothermia and the physiological effects of phosgene (Douglas, 2014). Or the involuntary use of patients at the Brooklyn Chronic Disease Hospital in which 22 patients were injected with live cancer cells (Douglas, 2014: 166). These are examples where information was knowingly withheld from research subjects which put them in direct harm.

Since the public became aware of unethical studies like the Tuskegee Study of Untreated Syphilis and other programs, ethics codes have been established to protect human subjects. In 1946, a group of German physicians who conducted medical experimentation in concentration camps were tried at Nuremberg during a 140-day long trial (Douglas, 2014). The trial resulted in the Nuremberg Code, a set of research ethics that influence modern concepts of informed consent and medical experimentation (Douglas, 2014). Wider public knowledge about the Tuskegee Study of Untreated Syphilis led to major changes in U.S law and regulation, such as the establishment of Institutional Review Boards for the protection of human subjects in studies involving them in 1974 and later the establishment of the Belmont report in 1978. The Belmont Report is one of the leading works concerning ethics and health care research and aims to ensure three principles: beneficence, justice, and respect for persons. Establishing the ethics codes aims to prevent future cases of misuse, with a particular focus on protecting human subjects, especially in clinical and other forms of medical research. In doing so, ethical values safeguard future risk to human populations.

Whenever there is a judgment of whether to proceed with an experiment, a decision must be made to decide what is more important: the protection of human subjects or the value of the knowledge to be gained (Douglas, 2014: 169). Since the value we place on human subjects often far outweighs any gain to knowledge, these studies can be halted on moral grounds. Per Douglas, “we can, as a society, allow the value of a particular piece of research to be outweighed by societal values” (Douglas, 2013: 7). Scientists can restrict the use of potentially harmful methods if the risk to human populations is too great. The central argument is that scientists must weigh the value of the research to be gained against non-epistemic values (Douglas, 2014).

An example of a societal value is the importance we place on human protection, another the importance we place on transparency. Within the literature there is another set of important questions, including ‘whose values are important’, and ‘who gets to decide’. While these are central questions, I will not address them here. The purpose of this introduction is only to set up the distinction between epistemic and non-epistemic values and offer examples of the role of non-epistemic values in scientific reasoning.

Ethics codes are constructed in narrow terms of concern for human subjects and rarely offer guidance about specific cases. There are many cases where ethical codes are unable to reduce the burden of value judgment on the scientist (Zahle, 2018), as I will go on to demonstrate.

2.1.2 Decisions about knowledge dissemination and use

If we follow Douglas, then a scientist cannot avoid their moral responsibility to consider all kinds of scientific activities, including decisions about knowledge dissemination about use. Think about the choice of whether to publish scientific findings, as Douglas does (Douglas, 2014). What is important for Douglas is that any competing concerns are weighed against one another to make a choice of whether to publish the findings. The scientist ought to consider any unintended consequences that the act of publishing might invite and weigh them against the value of publishing the results. It is in this way that a scientist can appropriately exercise their moral responsibility (Douglas, 2013).

Consider this with an illustration. There may be a situation where a social scientist conducts research into the causes of violence and finds that violence could be prevented by isolating different ethnic minorities in a country from one another by mass displacement (Cartwright and Runhardt, 2014: 282). In such a case, the scientist ought to question whether the action of publishing the result would be right given the possible consequences of doing so (supposing that there was strong reason to think that the displacement could and would then be implemented), as per Douglas (Cartwright and Runhardt, 2014: 282).

Scientists are often able to understand whether unintended consequences (or side-effects) might occur. We can imagine a scenario where a group of scientists develop expertise on the topic of violence mitigation, putting them in a better position to be able to understand the implications of publishing such a finding. One unintended consequence, as identified by the scientists, might be that the mass displacement of communities creates further ethnic divisions within the country. Other unintended consequences, identified by the scientists, might be that the isolation of ethnic minorities in a country nearly always further exacerbates different forms of violence, such as violence within ethnic groups rather than between them. Following Douglas, whenever a scientist can foresee the unintended consequences of publishing a result then they ought to weigh up whether publishing the result is the right thing to do, given those consequences (Douglas, 2014). So, the potential for further ethnic division in a community (known to scientists who are experts in this area) ought to be considered when deciding whether to publish the results.

Douglas' claim is that scientists have a moral responsibility to consider any side-effects of publication (Douglas, 2014). Douglas allows restrictions on publication where societal values are used to outweigh the value of producing or disseminating knowledge.

Importantly, Douglas does not say that scientists ought to consider the *benefits* of making a scientific choice. Douglas does not want scientists to act in support of political interests that they favour. Douglas warns that allowing scientists to use value judgment mustn't motivate scientists (or other interests) to make claims that favour one political outcome over another. That is, a scientist cannot allow their personal judgments about preferred violence mitigation strategies, for example, to influence the decision of whether to publish scientific findings. While, at times, it is appropriate to halt scientific publication on the basis that the results might cause foreseeable harm, the possibility for genuine discovery must remain on the table (Douglas, 2014).

Amartya Sen (1981) offers a related argument. Consider a paradigm case as outlined by Amartya Sen[2]. A BBC Panorama programme aired in 1980 cast doubt on the death of allegedly dead patients when their kidneys were removed, as recalled by Sen (1981). Choosing to air the BBC programme could have led to several unintended consequences: (1) the declining of organ donations; or (2) the increase in waiting time for those awaiting transplant. But choosing not to air the programme could have led to other unintended consequences: withholding important information from the public with a right to such information.

Should the BBC Panorama programme have released details of research on organ failure? There is an epistemic question: are the claims underpinning the BBC programme well supported by the evidence? and there is a moral question: is the action of broadcasting the right thing to do? Following Sen, a proper evaluation needs to consider any consequences of publication. It is insufficient to only assess whether the BBC programme claims are well supported by the evidence. For example, the claims aired on the BBC Panorama could negatively affect British healthcare practices. Or could influence citizens not to donate kidneys, thereby limiting the supply of available treatment. We may want to avoid these consequences to preserve the life-saving treatments that kidney donations offer. Sen argues that it can be possible for an account of kidney death to be scientifically accurate and well supported by the evidence. While at the same time, the act of publication may lead to undesirable consequences for society. We need both types of assessment to properly evaluate whether the BBC ought to publish the information on kidney deaths.

---

[2] See his paper from 1981 ‘account, actions, and values’.

Whether the BBC programme offers an accurate account can only be judged against the purpose for which the information is being sought (Sen, 1981: 100). If the purpose is to promote openness and public dialogue about the experiences associated with kidney death, then the BBC offers a good account. If the purpose, however, is to protect the lives of those waiting for transplant, then the BBC does not offer a good account (Sen, 1981: 100). Given the different value purposes, the BBC programme performs better or worse.

Consider this with an example. Think about the production of a policy statement of the mean consumption of a basket of goods as consumed by a British consumer (Sen, 1981). If the statement is intended to be used to describe the variety and amounts of goods consumed, then this is a 'good account', following Sen (1981: 98). But, if the statement is intended to offer an account of the standard of living in the UK, this is a 'bad account'. This is because the basket of goods is blind to distribution (Sen, 1981: 98). Following Sen, if the basket of goods consumed by a British consumer is blind to issues of distribution, then this could have a moral knock-on effect of making issues of distribution between groups invisible. Which could lead to the consequence of low-income groups being less well represented in policy. Hence, we can only judge the quality of policy statements (or publications) against the purpose that it is intended to serve (Sen, 1981: 97).

2.2 The "internal model"

It has been much harder to defend the use of non-epistemic (e.g., moral and political) values in the justification of scientific findings (Douglas, 2000). The argument from inductive risk is that non-epistemic values can influence the procedures for accepting a scientific conclusion when choice is underdetermined by evidence and there are risks to welfare from mistaken choice (Rudner, 1953: 2 in Douglas, 2000). Douglas (2000) claims that scientists can use value judgments in the very core of scientific investigation:

a) to consider whether the evidence is sufficient to warrant a particular claim given the consequences of wrongly accepting a claim versus wrongly rejecting it

b) in setting methodological standards like, for example, the level of statistical significance
c) in the categorisation of evidence
d) in the interpretation of results.

My discussion of a), b), c) and d) is taken from Douglas' 2000 paper, 'Inductive Risk and Values in Science'. I will draw on relevant illustrations from wider examples to further illustrate her arguments in this paper.

2.2.1 a) Accepting or rejecting a scientific hypothesis

Carl G. Hempel, C. West Churchland and Richard Rudner put forward the view that the scientist as scientist accepts or rejects hypotheses. According to this argument, scientists are forced to make value judgments when choosing the standards of evidence required for accepting a hypothesis (Churchman, 1948; Rudner, 1953 in Douglas, 2000). What counts as sufficient evidence for accepting or rejecting a hypothesis is "a function of the importance of making, in a typical ethical sense, a mistake in accepting or rejecting a hypothesis" (Rudner, 1953: 2 in Douglas, 2000). According to Rudner (Rudner, 1953: 2 in Douglas, 2000),

> Since no hypothesis is ever completely verified, in accepting a hypothesis the scientist must make the decision that the evidence is sufficiently strong or that the probability is sufficiently high to warrant the acceptance of the hypothesis.

This statement implies that sufficiency is a decision on the part of the scientist and should depend on the importance of getting the decision right or wrong. This has been taken to challenge the view that scientists can gather evidence or assess a hypothesis on the basis of the evidence without making value judgments. Rudner's famous example is that a scientist demands a relatively high degree of confirmation that pharmaceutical drugs are not toxic before accepting a hypothesis about the use of pharmaceutical drugs as treatment. A relatively high degree of confirmation is needed because the consequences of being wrong is very high according to moral standards. Put another way, when dealing with lethal toxins, it is typically more important to avoid false negatives than false positives. The issue of *how much evidence* is sufficient to accept or reject a hypothesis is just one decision point within the justification process in which there is inductive risk. As we will go on to see, Douglas points out that scientists face similar choices throughout the scientific process, and not only at the hypothesis testing stage.

As I noted, the argument from inductive risk as laid out by Richard Rudner (1953) and Carl. G. Hempel (1965) argues that the scientist as scientist accepts or rejects hypotheses (Douglas, 2000: 561). Because no evidence can establish a hypothesis with certainty “acceptance of a hypothesis carries with it the inductive risk that the hypothesis may turn out to be incorrect” (Hempel, 1965 in Douglas, 2000: 561). Douglas uses Hempel’s account to show that there are four possible outcomes of hypothesis testing,

> (1) ‘the hypothesis is accepted as true, and is in fact true
> (2) the hypothesis is rejected as false, and is in fact false
> (3) the hypothesis is accepted, but is in fact false; and
> (4) the hypothesis is rejected, but is in fact true’
>
> (Hempel, 1965: 92 in Douglas, 2000).

Of course, numbers three (e.g., false positives) and four (e.g., false negatives) are the types of error to be avoided.

Both Hempel and Douglas claim that the decision of whether to accept or reject a hypothesis depends on how the scientist – or perhaps the wider society in which the science is embedded -- values the various outcomes. There is a role for values when considering the possible outcomes, or “special instances of decision rules, such as maximising expected utility” (Hempel in Douglas, 2000: 562), where decision rules are rules for deciding or rules about accepting a hypothesis. Hempel provides a set of potential epistemic values to determine how one might consider the possible outcomes: “reliability, extensiveness, and systematization” (Hempel, 1965: 93 in Douglas, 2000: 562). These values are needed to weigh the consequences of the possible error that a scientist might make in accepting or rejecting a hypothesis (Douglas, 2000: 562).

Attempts to undermine Rudner's account have rejected the premise that the scientist as scientist accepts or rejects a hypothesis (Jeffrey, 1956). Jeffrey argues that scientists only assign probabilities. Scientists assign prior probabilities to the likelihood of the hypothesis and then update these theoretical priors when new evidence arises. This is the Bayesian process of conditionalization. According to Jeffrey, values play no logical role in the evaluation of epistemic evidence. The value judgments attached to the various outcomes are not the concern of scientists and scientists should restrict themselves to gathering and interpreting (see McMullin, 1983 in Douglas, 2000: 562).

Others also challenge Rudner's account. Levi accepts that scientists make judgments in their scientific activity but argues that these judgments should only be informed by "epistemic cognitive values" (1986: 43 in Douglas, 2000). Levi takes these to be indicators of truth (1986: 43 in Douglas, 2000). They might include predictive accuracy, fit with the data, simplicity, unification, and explanatory power. Levi does not support the use of other sorts of value influences to weigh upon scientific inferences (1986: 43 in Douglas, 2000). This set of strict epistemic value principles limits the type of value judgments scientists should make in carrying out their work[3].

Although this is not how she herself couches it, Douglas casts the acceptance or rejection of claims by the scientific community into the same framework as other actions that can be judged morally or practically. These are what Amartya Sen calls "ordinary actions" that can be taken by scientists (or not) and whether they should be is a matter of judging the consequences of error (Sen, 1981).

2.2.2 b) Setting methodological standards

Consider inductive risk in setting the appropriate level for statistical significance in studies of toxicology, as Douglas does (Douglas, 2000). Toxicology studies seek to evaluate the risk of

---

[3] This section will not cover other serious challenges to the argument from inductive risk. This is not the purpose of this chapter.

highly toxic compounds that are produced as by-products in some manufacturing processes, for instance herbicide production and paper bleaching. When evaluating whether a substance is dangerous, Douglas argues that we ought to think about the appropriate level of error that we are willing to accept (Douglas, 2000). We cannot do this without appealing to value judgments, or the interests that we want to protect. An excess of false positives about the harmfulness of dioxins means dioxins appear to cause more harm than they do. This would result in stronger regulation of chemicals and may in fact lead to the over regulation of chemicals. A harmful consequence would be the discouragement of investment in the industry. An excess of false negatives about their harmfulness, on the other hand, would mean dioxins appear less harmful and lead to a reduction in the regulation of unsafe chemicals. A harmful consequence would be the removal of restrictions on unsafe chemicals. This is essentially a balance between the risk of over-regulation or under-regulation and the evaluation of this risk ultimately depends on what we value. For example, while over-regulation would present excess costs to the industries that would bear the costs of regulations, under-regulation would present costs to public health, as Douglas argues (2000).

To take one of Douglas' renowned examples, cancer studies involving animals have been used to establish whether animals exposed to the harmful dioxins exhibit significantly more cancer than the control animals. Setting methodological standards is about deciding whether there is an optimum level at which the dioxins represent a safe dose. Again, this is a balance between false positives and false negatives. Douglas argues that a significant part of this decision is not epistemic but is about considering the consequences of the false positives and the false negatives. Where the consequences of the scientist's choice have clear non-epistemic consequences, such as the increase in risk to the general population, we need non-epistemic values in decision-making (Douglas, 2000).

Douglas says that values *should* enter in setting the methodological standards. Setting the methodological standards is essentially a balance between the risk of over-regulation or under-regulation; a significant part of this decision is not epistemic but is about considering the consequences of the false positives and the false negatives. Scientists ought to "consider the potential social and ethical consequences of error in their work, they should weigh the importance of those consequences, and they should set burdens of proof accordingly" (Douglas, 2009: 87).

Douglas counters the idea that discipline specific standards (e.g., 5% significance) determine the criteria for accepting or rejecting scientific claims[4]. She does not agree that when these standards are met, the hypotheses are appropriate to be accepted. To the contrary she argues that there is an important decision to make about whether to accept or reject the hypothesis and this decision depends on value consequences.

For example, in some instances we need to be concerned with the 1% when applied across a human population of millions (Douglas, 2000). In such cases, a scientist may appeal to notions of morality to determine whether the 95% significance rate is high enough for us to accept the hypothesis. The 5% may be a handy rule in some cases, but in others we might want to be much more cautious because of the very negative consequences of erring in one direction rather than another. Others claim that mistaken beliefs about p values cause much harm (Gigerenzer and Marewski, 2014). According to Gigerenzer and Marewski, the 5% standard is arbitrary, and we ought often to be concerned with the 1% and so demand a higher standard of evidence. Gigerenzer and Marewski claim that the automatic use of a statistical procedure over an informed judgment about the consequences of accepting a hypothesis incorrectly comes with risks.

2.2.3 An example of (b): Child safety standards

Safety assessments in child protection are used to predict the degree to which a child is likely to suffer maltreatment in the future. This type of assessment involves the systematic collection of information on threatening family conditions and current, significant, and clearly observable threats to the safety of a child. Social workers use a safety assessment as a measure to guide intervention. Eileen Munro considers the risk of error in the use of safety

---

[4] 'Significance level' refers to the threshold for statistical significance, which is often set at 5%. This is the chance of finding this association, or a stronger one, by chance, if none exists. So, the widely accepted choice of level for a statistical significance is 95%. This means that the false positive rate will be 5% or less.

assessments as a key professional activity in child welfare (Department of Health, 1995; Parton et al. 1997 in Munro 1999).

When evaluating whether a household is dangerous, we are thinking about the appropriate level of error that we are willing to accept. In practice, this is a balance of competing errors between false positives and false negatives. A false positive might lead to the outcome where a family being falsely accused of abuse when they are in fact innocent. Or a false positive might lead to the outcome of disrupting an innocent family, or the possibility of a child removed from home suffering abuse or harm in the alternative care situation provided (Munro, 1999: 123). While a false negative might lead to an outcome where a family is incorrectly cleared from an accusation of abuse when they are in fact guilty, which might lead to even greater harm for the child. Munro argues that the social work profession cannot avoid the balancing of these types of errors, which ultimately inform whether to intervene in the life of a child and family (Munro, 1999). A related issue, as outlined by Munro, is where a particular child is found with bruises of a certain sort. One social worker may not take these as evidence of abuse because of her background assumption that the child regularly does gymnastics and these bruises are characteristic of gymnastic injuries; whereas a different social worker may not believe that auxiliary, thinking the child and parent only *say* she does gymnastics regularly. What makes different putative facts evidence of abuse depends on background assumptions.

What is problematic is the disagreement about the weight of the various errors (Munro, 1999: 119-120). Social workers weigh the positive value of avoiding an outcome against the negative value of other outcomes differently (Munro, 1999). Munro argues that social work professionals are more aware than the public of the risks of taking a child into care (Munro, 1999). They might have acquired a unique experience of the emotional trauma for the child involved in rehoming a child in an unfamiliar setting. Whereas the public might much more readily want to avoid a situation where a child is exposed to the type of harm that they have seen in the media. For example, major trigger cases include the death of Maria Colwell in 1973 and later Victoria Climbie in 2000 and Peter Connelly in 2007. Maria Colwell was an English child who was killed by her stepfather in January 1973. The case was widely reported, and 40 years after the death the case remains in collective memory and is referred to when similar cases come to light.

The point at which professionals should act (e.g., "the threshold for intervention") is a value judgment that is made by policymakers (Munro, 1999). One option is for policymakers to respond to public pressures by lowering "the threshold for intervention" to minimise the chance of another child being subject to extreme danger, as occurred with the serious case of Maria Colwell. But this inevitably leads to more families with lower actual levels of abuse being caught up in the net (Munro, 1999: 124). A lower threshold for intervention also produces a higher rate of false positives (e.g., falsely accusing the families of abuse) and a higher threshold for intervention produces a higher rate of false negatives (e.g., missing cases of serious abuse). So, if policymakers set the threshold too high for intervention, then social workers might not intervene fast enough. But if policymakers lower the threshold for intervention, then social workers might be prevented from intervening in vulnerable households.

Munro puts forward a dilemma between *supporting families* on one hand and *protecting children* on the other. This is about balancing the needs of protecting children versus keeping families together,

> When family support has priority, *the threshold for removing children rises* and more children will be left in a dangerous setting. A death then triggers a swing towards prioritising protecting children. *The threshold for removal falls*, more children come into care and then there is a backlash when the general public fear that professionals are getting too powerful and invading the privacy of the family too readily (Myers 1994; Munro 1999, emphasis added).

Official policy clearly states that more attention must be given to supporting families and less time spent on expensive investigations of allegations, most of which do not reveal serious abuse (Munro, 2005). There is no acknowledgement, however, that reducing the number of detailed investigations done will lead, in some cases, to decisions about the seriousness of abuse being based on less information and hence being more fallible (Munro, 2005). Current government policy is creating another, new, conflict for practitioners. On the one hand, the welfare of the child must be the prime concern. While at the same time, the government had at the time introduced a complex system of targets and performance indicators, and an accompanying pile of paperwork, and they told social services that they will be evaluated and judged in the light of this system (Munro, 2005). Following Munro, the cultural and political context of child protection in the UK has covert and overt messages about what is valued and disparaged (Munro, 2005).

When dealing with lethal toxins, as Douglas does, it is typically much more important to avoid false negatives than to avoid false positives. Similarly, when dealing with child safety assessments, it may seem equally important to avoid false negatives given the implications for the social worker and child of failing to identify harm. Munro warns that while avoiding false positives might seem beneficial, stricter standards of evidence are not always preferable. When assessing child safety, extremely high standards can be stifling. There are already serious implications for failing to follow the correct procedure for identifying harms (Department of Health, 1995; Parton et al. 1997 in Munro 1999). This contributes to a tendency for professionals to overestimate error because the consequences of underestimation are so severe (Munro, 1999: 119).

2.2.4 c) Categorising and gathering evidence

Douglas considers inductive risk in the categorisation of evidence. She uses the example of scientists with the tendency to overestimate or underestimate the number of cancerous lesions which has an impact on how dangerous the dioxins are thought to be. When selecting a methodological approach, several difficulties arise in the categorisation of evidence. Some of the difficulties are unforeseen and arise from ambiguities in the data sources. For example, in dioxin cancer studies, female rats are dosed for two years and then auto-biopsies are carried out to gather endpoint data. The tissue and organ samples are mounted on slides later to be evaluated by toxicologists. However, Douglas shows that different pathologists have the tendency to evaluate the female rat liver slides differently. She shows differences in the way that the different pathologists judge tumours as malignant and benign. This lack of agreement among expert pathologists regards a judgment about whether a tissue sample has a cancerous lesion (Douglas, 2000: 569-570).

One approach is for a pathologist to take all borderline cases and categorise them as non-cancerous; but this has implications for the way that the evidence is used. This approach will make sure that there are few false positives but may lead to more false negatives. The consequence would be that malignant slides are underestimated, which leads to an underestimation of risk. Other pathologists may take an alternative approach and categorise all borderline cases as malignancies. This approach reduces false negatives but will likely increase the rate of false positives. The consequence would be an increase in the appearance of risk and could lead to more severe regulation. So, the process of evaluating the slides carries a significant uncertainty as to whether a judgment is correct. This uncertainty brings an inductive risk and the need to evaluate the possible consequences of error. Douglas argues that the pathologists should be concerned with the trade-off of false positives and false negatives. In other words, “one must decide what amounts (or levels) and kinds of inductive risk are acceptable” (Douglas, 2000: 569).

Value judgments are also sometimes needed to set levels on *how much evidence* is needed before drawing an inference. This is very similar to setting thresholds for acceptance/rejection of hypotheses that I discussed at the start. When the expected negative effects of a wrong decision much outweigh the positive effects of a correct one, more evidence is required. Just as we may set higher thresholds for certainty in these cases, we in train will want more evidence to achieve those levels. Epistemic judgments are of course required to decide how likely various alternatives are, but value judgments are needed to decide what counts as a positive effect and what as a negative and how they weigh up.

Consider inductive risk in the categorisation of qualitative data. When categorising data from children, there is a judgment of what (or who) to include or exclude, and this leads to differences in the way that different social scientists categorise the data. Scientists may operate with the tendency to overestimate or underestimate the importance of data from children with disabilities for example, which has an impact on how we understand the appropriateness of various programmes for these children. One approach is to treat all data as equal. Other scientists may take an alternative approach and decide not to use cases involving children with disabilities. So, scientists can either overestimate or underestimate the importance of the data because of underlying assumptions about the capabilities of children with disabilities. The consequence would be that children with disabilities are not as well represented in the dataset. The process of evaluating the data carries a significant uncertainty as to whether a judgment is correct. This uncertainty brings an inductive risk. Following Douglas, scientists need to evaluate the possible consequences of error, which might include the false characterisation of children of disabilities. Whether or not this risk is acceptable ought to be assessed.

2.2.5 d) The interpretation of results

Finally, Douglas considers the risk of error in the interpretation of results. She considers whether there is a threshold for dioxins' carcinogenic effects (Douglas, 2000: 575-577). Whether there is a threshold for cancer is affected by the parts of the evidence that the scientist chooses to emphasise and the background assumptions that the scientist appeals to[5] (Douglas, 2000: 575-577). This leads to different interpretations about whether there is a threshold. The cancer potency for most carcinogens (e.g., a substance capable of causing cancer living in tissue) can be assessed using a non-threshold model known as a "linear extrapolation model" (Douglas, 2000: 575-577). The linear extrapolation model implies that for any level of exposure there is an excess cancer risk associated with the exposure to the substance, and that risk will increase when exposure increases. In other cases, and for some substances, there is evidence that there may be a threshold for the carcinogenic effects. In such cases, what is known as a "threshold model" can be developed (Douglas, 2000: 575-577). A threshold model differs from a linear model because a critical dose point of departure is identified which can then be considered as protective of carcinogenic effects (Douglas, 2000: 575-577).

Douglas argues that even when there is supporting evidence that a substance may be a threshold carcinogen, the level of scientific uncertainty cannot lead to an obvious conclusion

---

[5] The argument from underdetermination will be covered in detail in chapter two.

(Douglas, 2000: 575). Different organisations and evaluators may disagree on whether a substance should be assessed using a threshold model or a linear extrapolation model, and consequently the organisations may derive very different guidelines. Whether or not the data shows evidence that there is a threshold for the dioxins to produce cancer is an interpretation. Thus, "[d]epending on which aspects of the data one chooses to emphasise and which assumptions one adopts, different interpretations arise" (Douglas, 2000: 575). These interpretations are morally significant.

As an illustration, as Douglas does, assume we adopt a threshold model and are wrong. This means that the regulations will likely be insufficient for public health. But if we adopt a linear extrapolation model (e.g., a non-threshold model) and are wrong, then the regulations will be overly severe. How the scientists and the community in which the science is done value these two possible errors should play an important role in assessing the inductive risk one is willing to take and which assumptions one is willing to adopt (Douglas, 2000). So, given a value-neutral assessment that the evidence supports the hypothesis to a certain amount, we must make a value-driven decision about whether that amount is high enough for us to accept the hypothesis (Douglas, 2000).

2.2.6 An example of d): The interpretation of results in climate models

Climate models also contain the risk of error in the interpretation of results. Climate models use quantitative methods to simulate the interaction of important drivers of climate change. They serve a variety of purposes, one of which is to make predictions of the extent of harm to people of climate change (Parker, 2014: 37). One prominent source of contention in the discussion of climate models is the "discount rate" for the welfare of future generations, as Parker discusses (Parker, 2014: 37). The discount rate reflects a choice that is made by scientists that is critical to the assessment of how great the harms are overall, considering now and in the future. To set the discount rate at an optimal level the scientist must appeal to value considerations. A discount rate is "a parameter (or group of parameters) that represents the extent to which goods in the future are considered less valuable than similar goods obtained today" (Parker, 2014: 42). A discount rate essentially serves as a way of weighing the interests of people at different times, e.g., current versus future generations.

Where to set the discount rate has been much disputed in the field of climate science. It is not merely a disagreement about empirical matters but the different approaches to the choice of discount rate suggests an appropriate role for ethical considerations in economic analysis (Parker, 2014: 42).

A reliance on different discount rates led two economists to reach opposing conclusions on mitigation policies. Nicolas Stern (a British economist) conducted the Stern Review on the Economics of Climate Change and employed a discount rate of 1.4 percent (Stern 2006). This was later used to support a policy of reducing greenhouse gas emissions by approximately 3 percent per annum. Stern concluded that “even mitigation policies requiring significant near-term emission reductions would have benefits that far exceed their costs” (Parker, 2014: 42).

However, William Nordhaus (an American economist) argues that the discount rate was too low and thus challenged the basis of Stern’s conclusions. Nordhaus argues that a higher discount rate of 6 per cent has greater justification (Gouldner and Williams, 2012). The choice of discount rate could be based almost entirely on ethical considerations, with no need to ground the discount rate in observed or expected estimates of the opportunity costs, as Stern argued: the welfare of each generation is equally valuable (Stern in [6]Parker, 2014). But, in contrast basing the choice of discount rate on observed behaviour, such as behaviour that is reflected in market interest rates has greater justification according to Nordhaus (Parker, 2014). The significance of Norhaus’ higher discount rate, according to Gouldner and Williams (2012), is that it tends to shrink the present value of benefits relative to the present costs and weakens the case for aggressive current action; this means that relatively small differences in the choice of this rate can make very large differences in the policy assessment (Gouldner and Williams, 2012).

If the scientist is optimistic about the risk to future societies, like William Nordhaus, then the discount rate may be set at a higher rate. If the scientist is concerned about the risk to future societies, like Nicolas Stern, then the discount rate must be set at a lower rate. So, if we care

---

[6] In principle this would mean no discounting at all future welfare. Stern didn’t do it because he also used the discount rate to represent various uncertainties in his model.

heavily about future generations then we can set the discount rate one way. Nordhaus, though, takes a different ethical stance. By contrast with Douglas, he supposes that the scientist should not be imposing their own value judgments on a problem. He instead looks to the average behaviour of real people now to reflect the values that should be used, their willingness to sacrifice their current welfare for the future welfare of themselves and their families, as reflected in the savings rate. I want to point out that whether we let Stern and other cross-generation egalitarians set the discount rate or let current savings practices do so, this is in fact an ethical decision about how such matters should be resolved. "Let current people decide" is no more value-free than "Treat all generations equally". No matter which way, this is not a value-free or value-neutral decision. We can say that any choice of discount rate requires moral judgments.

2.2.7 A second example of d): Application of scientific findings

Another area with possible error is where scientists overestimate the effects of a drug trial, for example, making it appear as though claims travel further than they actually do. This is an additional decision point at which there is a risk of error.

Consider the worldwide roll out of the vaccination for Human Papillomavirus (HPV), as Intemann and de Merlo-Martin do (Intemann and de Melo-Martin, 2010). There are over 100 different strains of the HPV infection, varying from "high-risk" to "low-risk" according to the likelihood of the infection turning cancerous. HPV is prevalent worldwide, but around 83% of cancers occur in Latin America, sub-Saharan Africa and South and South-East Asia (Parkin et al. 2005 in Intemann and de Melo-Martin, 2010). Gardasil and Cervarix were recently approved as vaccinations to protect women against infection from the high-risk types of HPV. Gardasil protects against most cervical cancers in the UK (e.g., what is known as types 16 and 18[7]). The National Health Service in the UK, for example, uses the vaccine Gardasil for its countrywide vaccination programme.

---

[7] Two HPV types (16 and 18) cause 70% of cervical cancers and pre-cancerous cervical lesions.

When it comes to evaluating the effectiveness of the vaccine, there are additional standards that are necessary before we can claim that the results from one setting transfer to another. Gardasil and Cervarix, according to Intemann and de Merlo-Martin, are more suited to women in developed countries like the UK and so claims about their effectiveness worldwide may carry an inductive risk. They claim that tests on women in developed countries are poor measures of effectiveness for women elsewhere. In practice, the implementation of vaccination programmes is important. Screening programmes are widespread in the UK and the USA, but they are less prevalent in parts of Asia or Africa. So too is the health service better available in the UK and USA than in other countries. There are also cultural customs that likely decrease young females being able to enrol and stay in a vaccination programme that requires up to three doses of the same vaccine. To be successful in different population contexts, the vaccine must account for the circumstances in which the vaccine is to be distributed, as Intemann and de Merlo-Martin point out. One important choice is whether to prioritise evidence of long-term effects of the HPV vaccine in developed settings or the value of getting the vaccine quickly to different markets. Making a good decision requires value judgments (Intemann and de Melo-Martin, 2010). Value judgments play an important role in our standards of evidence by helping us to determine how confident we can be that a programme is having a beneficial impact given a particular goal.

If the goal is to decrease instances of mortality and morbidity worldwide, then a vaccine that will work in developing countries is essential too, and our inclusion criteria for trials on the vaccines look very different[8] (Intemann and de Merlo-Martin, 2010: 3). If the goal is to focus on industrialised countries, or only a specific strain of HPV relevant to only a certain part of the population, then our inclusion criteria would be much more distinct. If we care about an average reduction worldwide then this would require targeting the vaccination programmes to African countries where the provenance of the HPV virus is much more dominant, for example (Intemann and de Melo-Martin, 2010). Alternative evidentiary standards are necessary if the vaccination programme is to be useful in different types of settings (Intemann and De Merlo-Martin, 2010: 2).

---

[8] Consider other examples as outlined in other research from Intemann on women's fertility drugs, for example.

2.3 Summary

Douglas distinguishes between a direct and an indirect role for values in scientific investigation (Douglas, 2000). She argues that social and other non-epistemic values (e.g., moral and political) may indeed play a direct role in various external aspects of scientific research – such as setting the question for study, restricting the methods of study and publishing results. Both epistemic and non-epistemic values can always be used in the external stages of science. However, non-epistemic values only have an indirect role in the internal stages of science: a hypothesis must not be accepted because it fits a scientist's agenda. This would be a direct role for values in the internal stages of science, which Douglas does not permit. If we follow Douglas, though, non-epistemic values must enter at external stages, and especially in deciding how certain we should be before accepting or rejecting a hypothesis. It is morally irresponsible not to make reasoned value judgments weighting harms and benefits where producers of knowledge are in a position to do so.

Social and other non-epistemic values may be invoked to require more evidence, or to tolerate less uncertainty, when accepting a theory. In this sense, social and other non-epistemic values play an indirect role, by modulating the evidential threshold, but not playing the same kind of role as evidence. For Douglas the only way that it is appropriate to use values at any stage in evaluating the acceptability of a hypothesis is in assessing the *inductive risk* of one procedure over another and using that assessment to decide which to choose. Douglas does not want values to enter in deciding whether evidence *is* evidence for a hypothesis, nor in deciding how strongly the evidence supports the hypothesis. For her, value judgments should not directly influence the logical role of accepting or rejecting a scientific claim but inform whether the evidence is sufficient to warrant a particular claim given the consequences of wrongly accepting a claim (Douglas, 2000).

Douglas is talking about *scientists'* needing to make value judgments. More broadly this can be taken to defend the claim that: *the job of science cannot be properly done without value judgments being made*. But by whom? Scientists? Or some external body, for instance an Ethics Review Board, or a pharmaceutical company in cases where the scientist has signed away the right to publish sponsored research? There can be an issue then of who 'should' make the value judgment. Douglas stresses role of scientists as important since they are often in a position to recognise well many of the possible effects. But she does not really go into who should make the value judgment. Nor do I. My purpose is only to point out that value judgments are inevitable – choices need to be made and they will have value consequences regardless of whether anyone deliberates about which should dominate. The other issue Douglas does not take on is whose values get used. The scientist could, for example, decide in accord with their own values that don't match 'society's' and different bits of society may have different values. But again, this is not an issue that I will tackle in this introductory chapter.

What does this mean for the policies I describe?

The process of systematically collecting facts to evaluate whether a household is dangerous or not requires the use of values. For example, what makes facts evidence of child abuse depends on what further assumptions are made to turn those facts into claims bear on the truth or falsity of an abuse claim.

Similarly, in the climate modelling example, the choice of where to set the discount rate depends on the background assumptions about the importance to be attributed to the utility of later generations relative to earlier ones. Hence, values (encoded in background assumptions) play a pervasive role in the production of evidence to inform (a) child safety assessments; and (b) assessments of harm in climate modelling. In subsequent chapters, I will go on to explore the role of values in other areas of social policy.

2.4 Conclusion

Douglas allows for the inclusion of social, political, cultural, and ethical values in evidential assessment of hypotheses through an assessment of risks to humans and other things we value. Rather than limit the kinds of value that can legitimately play any role at different stages of research, we should limit the *kinds of role* that any value can legitimately play at different stages of research. The value-laden ideal gives diverse values a necessary and legitimate role throughout the research process, which is direct at early stages (e.g., restrict harmful data) and indirect at later ones (e.g., hypothesis testing).

References

Cartwright, N. and Runhardt, R. (2014) Measurement, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 31-47.

Churchman, C. W. (1948) Statistics, Pragmatics, and Induction, *Philosophy of Science*, 15: 249-268.

Clements, M. and Hendry, D. (1999) On winning forecasting competitions in economics. Spanish Economic Review (1), 123-160.

Department of Health (1995) *Child Protection: Messages from Research*, London: HMSO.

Douglas, H. (2000) Inductive Risk and Values in Science, *Philosophy of Science*, 67, (4), 559-579.

Douglas, H. (2008) The Role of Values in Expert Reasoning, *Public Affairs Quarterly*, Vol. 22, No. 1, pp. 1-18.

Douglas, H. (2014) *Values in Social Science*, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 162-185.

Elliot, K. (2013) Douglas on values: From indirect roles to multiple goals, Studies in History and Philosophy of Science, 44, 375-383.

Efstathiou, S. and Mirmalek, Z. (2014) Interdisciplinarity in Action, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 233-248.

Gere, C. (2017) *Pain, Pleasure and the Greater Good: From the Panopticon to the Skinner Box and Beyond*, The University of Chicago Press: Chicago and London.

Gigerenzer, G. and Marewski, J. N. (2014) Surrogate Science: The Idol of a Universal Method for Scientific Inference, Journal of Management, https://doi.org/10.1177/0149206314547522

Gouldner, L. H. and Williams, R. C. (2012). The choice of discount rate for climate change policy evaluation, Climate Change Economics, 3, (4), 12 pages.

Hempel, C. G. (1965) Science and Human Values, *in Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* New York: The Free Press, 81-96.

Intemann, K. and de Merlo-Martin, I. (2010) Social values and scientific evidence: the case of the HPV vaccines, *Biology and philosophy*, 25, (2), 203-213.

Jeffrey, R. C. (1956) Valuation and Acceptance of Scientific Hypotheses, *Philosophy of Science*, 22: 237-246.

Levi, I. (1962) On the Seriousness of Mistakes, *Philosophy of Science*, 29: 47-65.

Longino, H. (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton, NJ: Princeton University Press.

McMullin, E. (1983) Values in Science, in Peter D. Asquith and Thomas Nickles (ed.), *Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association, Volume 1.* East Lansing: Philosophy of Science Association, 3-28.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harpercollins.

Munro, E. (1999) Protecting children in an anxious society, *Health, Risk & Society*, 1:1, 117-127, DOI: 10.1080/13698579908407011

Munro, E. (2005) Improving practice: child protection as a systems problem, *Children and Youth Services Review*, 27, (4), 375-391.

Myers, J. (1994) *The Backlash: Child Protection under Fire,* London: Sage.

Parker, W. (2014) Climate Change, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 31-47.

Quine, W. V. O., 1951, "Two Dogmas of Empiricism", Reprinted in *From a Logical Point of View*, 2nd Ed., Cambridge, MA: Harvard University Press, pp. 20–46.

Parton, N., Thorpe, D. & Wattam, C. (1997) *Child Protection: Risk and the Moral Order*, London: Macmillan.

Reiss, J. (2017) Fact-value entanglement in positive economics, Journal of Economic Methodology, 24, (2), 134-149. DOI: 10.1080/1350178X.2017.1309749.

Rudner, R. (1953) The Scientist Qua Scientist Makes Value Judgments, *Philosophy of Science*, 20: 1-6.

Sen, A. (1981) 'Accounts, Actions and Values: Objectivity in Social Science', in C. Lloyd (ed.), Social Theory and Political Practice: Wolfson College Lectures 1981. Oxford: Clarendon Press, 87-107.

Stern, N. (2006). *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

Stiglitz, J. (2009) The anatomy of murder: Who Killed America's economy? Critical Review, 21, 329-339.

Zahle, J. (2018) Values and Data Collection in Social Research, *Philosophy of Science*, 85, (1), 144-163.

Chapter 3: Underdetermination of theory by evidence

3.1 Introduction

This chapter looks at and defends a particular role for values. Helen Longino describes "contextual values" as values that can play in the process of scientific legitimation of hypotheses, which many have urged should be value-free (Longino, 1990: 216). I draw on the work of philosopher of science Helen Longino in particular. Longino has written extensively on the significance of contextual values to scientific inquiry (1979, 1990, 1992, 2002, 2014) and is a central figure in feminist epistemology. Longino defines contextual values are those that belong to the social and cultural environment where science is done and include the norms, preferences, beliefs, or interests about what is acceptable or unacceptable practice, and which vary from community to community (Longino, 1990). These may, and often do, include what we would normally judge morally relevant norms, preferences, etc – like the preference to serve one social group (perhaps oneself and one's family) over others. I defend a role for values in the collection and processing of evidence: in the selection and collection of facts, the rendering of facts as evidence and the weaving of these together to confirm or disconfirm a hypothesis.

I explain the arguments of contextual empiricist and feminist philosophers of science that there is no use in lamenting the use of contextual value in this process since the process cannot be caried out without it. It is inevitable whenever we engage in "ampliative inference". If the inference is ampliative—going beyond the observed facts—the hypothesis will be underdetermined by the facts, as the famous Duhem-Quine thesis teaches. In science this is done by background assumptions, and these will sometimes reflect contextual values, including moral, political, and cultural values and other points of view. Feminist standpoint theorists use this to argue for the importance of including minority researchers and the subjects of research in research teams since they have special epistemic access. Longino and others use it to argue for diversity of viewpoints. I don't side one way or another here. My chief concern is not to discuss whose values should be included but rather to make clear the argument that the use of contextual values is inevitable so the issues this raises must be faced up to[9].

3.1.1 Chapter outline

First, I outline the argument from underdetermination with some illustrations (e.g., Longino, 1979; Rolin, 2012; Di Bucchianico, 2009; Atkinson, 2001). Second, I offer examples from feminist empiricist philosophers who claim a role for contextual values in health science (e.g., Longino, 1990, 2002; Anderson, 2002, 2004; Crasnow, 2014). This includes efforts to study the concept of self-defence (Crasnow, 2014) and divorce (Anderson, 2004). Finally, I apply these ideas to a new example of the 2014 West African Ebola outbreak (e.g., Wilkinson and Leach, 2015; Fairhead, 2016). What I intend to illustrate with my example of the 2014 West Africa Ebola outbreak is how a commitment to a single value perspective can restrict how we understand the spread and containment of Ebola.

The key is that we leave ourselves open to surprise (Anderson, 1995). We need to ensure that value judgments do not operate to drive inquiry to a predetermined conclusion (Anderson, 1995). When values push scientists to corroborate already endorsed claims rather than impartially assess the evidence for and against them, this threatens the reliability of belief formation in science, as it contributes to the one-sided processing of facts (Anderson, 1995).

I consider feminist empiricist philosophers' use of contextual values in what I pick out as three distinct ways: in the selection and collection of facts, the rendering of facts as evidence and the weaving of these together to confirm or disconfirm a hypothesis. My first category refers to the collection of facts. This refers to the process of finding out what the facts are. The second category refers to the use of facts as evidence. Nothing comes labelled as evidence for a given hypothesis. There must be a reason to see the obtaining of given fact

---

[9] There are an additional set of concerns. These are concerns about the specific role of the scientist. How does the scientist make visible their own values? Should a scientist think and act critically? How well does the scientist accurately reflect the different values, beliefs, and interests of various groups? These are an important set of questions, but I will not address these concerns.

as evidence for the truth of some other claim. I then add a third category - the weaving of these together to confirm or disconfirm a hypothesis. My third category refers to the way that certain kinds of facts are emphasised in the production of evidence for a hypothesis.

3.2 Underdetermination in science

Underdetermination refers to the existence of a gap between evidence and hypothesis. Background assumptions, as I will go on to describe, fill in this gap to link evidence to hypotheses (Longino, 2002).

The most famous account of underdetermination is by Pierre Duhem (1956) and Willard Van Orman Quine (1951) and is widely known as the "Duhem-Quine thesis" or "holist underdetermination". This is the thesis that all scientific hypotheses, theories, and models are underdetermined by the evidence, where the data available could be insufficient to determine what beliefs to hold in response to it (Longino, 1990). There is nothing intrinsic to the data or its relationship to the hypothesis that establishes an evidential relationship between them (Longino, 1990). In the absence of formal connections, data acquire their status as evidence from background assumptions about the relevance of the data (or an observation) to the hypothesis or theory (Longino, 1990). To determine what belief to hold in response to the evidence, determinations are needed about the relevance of the data to the hypothesis under investigation and ultimately what data will be counted as evidence in support of the scientific hypothesis (Longino, 1990).

According to the 'holist view' of theory confirmation, auxiliary, or background assumptions, are necessary alongside a hypothesis (Duhem, 1962). Especially since theory could be taken broadly enough to include the requisite background assumptions in order to generate predictions. It is almost never the hypothesis alone that predicts the data and hence never the hypothesis alone that is confirmed or disconfirmed by the data (Duhem, 1962: 185). A scientist can either reject a hypothesis on the basis of evidence that seems to falsify it, or a scientist can decide to maintain that the hypothesis being tested could be accurate, even in light of evidence that seems to falsify it, by rejecting the auxiliary assumptions that she used to test the hypothesis (Duhem, 1962). Either way, the choice of which hypothesis to accept is not only an empirical decision based on data and facts alone but is made in virtue of background assumptions.

Consider some simple examples. One common case is where the evidence available to us does not tell us which hypothesis to accept over another. One simple example is in determining whether a child has measles (Rolin, 2012). When assessing whether the facts (e.g., a child's body is covered in red spots) fits the hypothesis that the child has measles, or whether the facts support an alternative hypothesis that the child has an allergic reaction, the facts can offer support for both hypotheses. Because of this challenge, background assumptions are necessary to establish the evidential relevance of the facts to the hypothesis. We can say that the facts more strongly support the hypothesis that a child has measles because we gather data about other symptoms (e.g., the child has a temperature, hacking cough, runny nose) and our background assumptions include the knowledge that these are highly symptomatic of the illness. Facts only become evidence given assumptions.

In fact, even without conflicting hypotheses, the facts won't imply the hypothesis and seldom even does a hypothesis imply the facts.[10] Neither "the child has measles" nor "the child has an allergic reaction" imply "a child's body is covered in red spots". If one wants to link fact statements as evidence to hypotheses, then one must use background information about the symptoms of measles, etc.

Longino puts this slightly differently. "If one believes that red spots are a symptom of the measles, then one will take the presence of red spots as evidence that she has the measles" (longino, 1979: 38). If "one believes not that the red spots are a symptom of the measles but that they are symptomatic of some gastric disorder [then] one will take the red-spotted stomach as evidence that she has some gastric ailment" (Longino, 1979: 38). What determines whether we take a fact (or alleged fact) as evidence for a hypothesis are the beliefs that a person holds concerning the evidential relation between the fact and hypothesis (Longino, 1979: 37). It is these beliefs which determine whether the fact is evidence of a hypothesis (Longino, 1979: 37).

---

[10] Here is a possible exception. Hypothesis: "All ravens are black." Factual claim: "If this is a raven, it is black". But this latter would not usually be thought to be evidence, in contrast to "This is black."

Consider an illustration where all actors have access to the same facts but disagree about what the facts tell us. Di Bucchianico offers a case study to illustrate two warring camps offering conflicting theories of high temperature superconductivity (Di Bucchianico, 2009). One camp took the explanatory mechanism to be *phonons*, while the other took the explanatory mechanism to be *magnetic modes, as* Di Bucchianico argues. With new procedures (known as Angle-resolved photoemission spectroscopy) that produce 3D images, a new fact was discovered. The new procedures demonstrated a ‘kink’ in the dispersion curve of reflected photoelectrons. Each side, because of the way they interpreted the kink, claimed both that the new fact was evidence that supported their theory and that it was inconsistent with the opposition’s theory. As Di Bucchianico argues, the two camps held widely different interpretations of it due to the great number of differing assumptions and practices they were also committed to (Di Bucchianico, 2009). We may be impressed with a new result and take it to confirm our hypothesis, but whether it does depend on a host of other assumptions being true, such as other experiments having been well conducted, other concepts being true to the world, and so forth[11].

What these examples show is that why someone takes a fact as evidence for one hypothesis over another depends on one’s other beliefs (Longino, 1979: 40). We can call these beliefs background assumptions (Longino, 1979: 40). Facts by themselves cannot direct a scientist to accept or reject a hypothesis. The choice of what counts as evidence for or against a theory is made in virtue of background assumptions. What Di Bucchianico points out, for example, is that one’s assumptions vary depending on the practices they are also committed to (Di Bucchianico, 2009). We may hold differing assumptions depending on our disciplinary background and/or other commitments. Much of the time scientists leave background assumptions implicit. But whether a scientist accepts a hypothesis depends on these assumptions. So, “to talk about something being evidence independently of its being taken so by some person is to introduce an implicit assessment of its value” (Longino, 1979: 36).

---

[11] This phenomenon can be seen in the construction of rival accounts that posit different unobservable entities that are responsible for the empirical consequences and is seen in cases where there is more than one explanation for the existence of the phenomenon. In such a case, it is always possible to construct an alternative to the leading theory that has exactly the same empirical consequences.

3.2.1 Examples from economic modelling

Consider the role of background assumptions in economic modelling. Economic models are based on several typically implicit assumptions which ought to be questioned. Atkinson famously argues that if one accepts a model and the evidence it offers then they are accepting the assumptions inherent in the model. One example is economists' frequent use of *a representative agent model*, which represents an entire group in the same way, as Atkinson argues (Atkinson, 2001). These models ignore and put out of focus issues about distribution within that group that are relevant to questions of justice. It may be assumed that changes in welfare can be judged by whether, for example, the representative household is better off or worse off. But when an economist adopts a representative agent model this can have the moral knock-on effect of making issues of distribution between individuals in the household invisible. In most policy situations there are conflicts of interests and so household heterogeneity is a crucial aspect to consider. Economists use background assumptions to decide between types of economic models to use. The assumption of a representative agent may suffice for studying some macroeconomic behaviours, but doesn’t focus on other problems (e.g., distribution[12]).

We see this in models that maximise expected utility[13] (Atkinson, 2001). Choosing to maximise expected utilities makes sense as a modelling strategy if one accepts utilitarianism as the right moral framework for those kinds of policy decisions. While it may be assumed that there is agreement on a utilitarian welfare criterion (e.g., maximising utilities) there are other alternatives to individual utility. This includes Rawls’ principle of basic liberties (Atkinson 1973 in Sandmo, 2017). Or a Rawlsian approach to distributional justice (Atkinson 1973 in Sandmo, 2017). A Rawlsian approach focuses on the minimum levels of welfare, which is contrasted with traditional summation of utilities (Atkinson, 2001: 202). Rawls was

---

[12] For researchers who are committed to the worst-off, attention to distribution is essential (Longino, 1994: 477 in Anderson, 2014: 17).

[13] Integrated Assessment Models used to investigate climate policy are also based on a number of typically implicit assumptions which ought to be questioned. One is the justification for discounting utility. Another is the justification for how we ought to weight different generations.

committed to the distribution of what he called primary goods as opposed to the distribution of individual welfare. One example is the "optimal marginal tax rate" which differs with a Rawlsian rather than the utilitarian criterion for social welfare (Atkinson 1973 in Sandmo, 2017). Within a Rawlsian approach we would maximise the utility of the least advantaged rather than maximise the sum of utilities. Hence, the optimal marginal rate of income tax might become considerably higher. Atkinson claims that these assumptions ought to be made explicit in our models and our economics (Atkinson, 2001: 196).

Finally, consider average treatment effects (ATE), as Deaton and Cartwright do. ATE is a measure used to compare treatments in randomised experiments, evaluation of policy interventions, and medical trials (Deaton and Cartwright, 2018). ATE corresponds to the average health programme effect for the entire population, regardless of whether or not a particular individual benefits. ATE does not therefore account for heterogeneity in treatment outcomes (Khosrowi, 2019). Deaton and Cartwright argue that choosing a method that is good for estimating an ATE over other methods that are better at getting information about individuals only makes sense if you *value* information for public health planning that can rightly be concerned with improving population averages (Deaton and Cartwright, 2018). If one values the welfare of a specific individual, a different method may be better.

These examples raise an important concern. What counts as *relevant* data to inform a model is guided by our values in deciding which data to represent or how to represent them. As soon as an economist settles on a model, we may end up squeezing out certain values and interests[14] (Atkinson, 2000). We often do not only want to know whether an intervention is effective on average but also whether the intervention is equitable, or how it affects different users[15] (Khosrowi, 2017). Distributional type information, for example, is particularly valuable if we are interested in improving the lives of the worst well off as well

---

[14] See Biddle and Kukla for a similar argument, that we may end up miscategorising data, undertreating people, pursuing certain questions over others and concealing interests (Biddle and Kukla, 2016: 229).

[15] Highlighting "heterogeneity" as a feminist goal allows feminists to represent diversity as a potential resource (Anderson, 1995: 31). It is because heterogeneity is such a valuable potential that it is important to include it in our conceptual schemes (Anderson, 1995: 50). This is not an argument that I go into.

as improving averages. Which model is most useful depends on the values that are to be served.

3.3 A role for contextual values

Longino famously argues that background assumptions are the means by which what she calls “contextual values” and ideology are incorporated into scientific inquiry (Longino, 1990: 216). Contextual values include the norms, preferences, beliefs, or interests about what is acceptable or unacceptable practice, and which vary from community to community (Longino, 1990). Scientists might, for example, be linked to preferences for the scientific institution or tradition they belong to or might be affected by a sense of loyalty to the scientific community to which they belong (Longino, 1990). Or alternatively, scientists may be affected by their worldviews, interests, and the special advantages afforded to theories in terms of their problem-solving abilities or because of other so-called ‘epistemic virtues’ like simplicity (Longino, 1990: 4; 81). Contextual values are often deeply held, and scientists cannot be aware of all the background assumptions they presuppose in accepting a hypothesis (Longino, 1990). It will be very difficult for individual scientists to recognise when their work is being influenced by their own contextual values and experience (Longino, 1990, 1992). As I will go on to argue, the use of contextual values can lead to the misrepresentation or dismissal of facts that for someone else, with different background assumptions, might count as evidence.

First consider some of Longino’s well known illustrations of individual scientists who express the contextual values inherent in their cultures and particular lives when they engage in a scientific activity (Longino, 2002: 135-136). We see this in the claims made by late nineteenth-century Europeans that the skulls and posture of European races were more developed than ‘Negroes’ (Gould, 1981 in Longino, 2002). Or, in claims made by male scientists that female skulls’ skeletal anatomy and physiology were evidence of women’s “natural” role in society (Allchin, 2015). Or with the familiar illustration of nineteenth century theories of intelligence biased by the widespread belief that women were inferior to men, which made it easier to explain away data suggesting similarities in cognitive abilities (Gould, 1996 in Intemann, 2010: 780). All of these cases were once considered good science from their own historical context (Allchin, 2015). What these cases are used to show is that all kinds of contextual values can and do enter science; but more importantly, that scientific conclusions often reflect the biases and values of its practitioners (Longino, 2002: 135-136). What is needed is further consideration of how this process occurs, which I will go onto provide with my example of the Ebola outbreak.

Longino and others argue that it is the diversity of viewpoints that matters. For example, “it is the diversity of the contextual values represented and not the content of any particular contextual value judgment that makes for good science” (Intemann, 2005: 1006). The idea is that it is easier to identify when values inappropriately influence scientific reasoning when those values are different to one’s own (Longino, 1990). So, scientific communities with diverse values provide a sort of check against biasing values and interests to ensure that they do not inappropriately influence scientific reasoning. For example,

> The evidential relevance of particular elements of experience to hypotheses is mediated by background assumptions operating at many levels. What controls the role of background assumptions is interaction among scientists, interaction consisting in criticism of assumptions involved in observation, of assumptions involved in reasoning, of assumptions involved in thinking a given hypothesis is plausible, of assumptions involved in the application of particular methods to the solution of particular problems. To be successful in uncovering such assumptions, criticism must proceed from a variety of points of view, ideally as many are as available (Longino, 1992: 334).

Feminist standpoint theorists, on the other hand, argue for the importance of including minority researchers and the subjects of research in research teams since they have special epistemic access (Intemann, 2005). For feminist standpoint theorists, it is the specific experience and not just the diversity of social positions that is epistemically beneficial (Intemann, 2005). For instance, a person’s particular experience is what is more likely to identify the limitations with background assumptions that have been unnoticed. I don’t side one way or another here. What is important for my purposes is that both sets of arguments commit to the view that science neither is, nor ought to be, neutral among social and political values (Intemann, 2010). My chief concern is not to discuss *whose* values should be included but rather to make clear the argument that the use of contextual values is inevitable.

Feminist empiricist philosophers claim that underdetermination leaves open the permanent possibility that our values may be covertly influencing our assessments of the evidence (Anderson, 2002). These philosophers contradict a central argument of mainstream empiricist philosophy of science: they reject the claim that scientific theories do not presuppose any non-epistemic value judgment (Anderson, 2002). Feminist empiricists further reject the claim that scientific theories do not serve any particular non-epistemic value more fully than others (Anderson, 2002).

For example, Elizabeth Anderson (1995, 2002) criticises mainstream science and technology on the following basis: science and technology does not serve women's interests; science and technology reinforces women's subordination; science and technology often represents women's subordination as something natural. What is alarming for feminist empiricists like Anderson is that widely accepted scientific theories may reflect androcentric or sexist values (Anderson, 1995). I will consider three examples of androcentrism or sexist values in science before turning to my main example of the Ebola outbreak.

3.3.1 Illustration 1: Gross national product

One way in which contextual values influence science is in the description or definition of phenomena from the perspective of men, or typically male experiences, without consideration of how the same phenomena might be described by women (Anderson, 1995: 71). Anderson argues that there is implicit androcentrism in standard economic definitions like gross national product (GNP) (Anderson, 1995: 71-72). Following Anderson (1995), economists adopt an androcentric conception of a household, which assumes that a household consists mainly of a male head of household and a female who is supported by the male in performing mainly childcare and other household duties (e.g., “housework”) (Anderson, 1995: 72). For example,

> The androcentrism implicit in the standard economic definition of productive labor has profound implications for national income accounting, the fundamental conceptual framework for defining and measuring what counts as economically relevant data for macroeconomic theory. It effectively excludes

> women's gender-typical unpaid domestic labor from gross national product (GNP) calculations, making women's work largely invisible in the economy (Anderson, 1995: 71-72).

This has particularly harmful implications for women in developing countries, who perform a majority share of typically unpaid domestic tasks. Economists may choose to assign a market value to various unpaid domestic tasks if they recognise that a significant proportion of productive labour is found in domestic production activities associated with subsistence agriculture and home construction (Anderson, 1995: 72). However, an economist must choose which of these tasks count as *productive labour* (Anderson, 1995: 72). The choice of which tasks count as productive labour is not value-neutral as it has a potentially harmful effect on women in policy. Anderson claims that economists often adopt an androcentric conception of a household, which assumes that a household consists mainly of a male head of household.

> [Economists] assume that households consist of a productive primary producer, the husband, who supports a wife engaged in "housework, "which is assumed to be economically unimportant or unproductive. "Housework" has no clear definition in societies where most production takes place within the household. So economists apply the concept of "housework" to whatever productive activities a society conventionally assigns to women. Thus, women's unmarketed labor in these societies counts as productive only if men usually perform it too, whereas men's unmarketed labor is usually counted in the national income statistics regardless of its relation to women's labor (Waring 1990, 74-87 in Anderson, 1005: 72).

There are numerous examples of androcentric bias within the literature (Kabeer, 2012; Chopra, 2014). For example, in Malawi and Lesotho, women play a majority role in the production of food for household consumption, yet policy interventions have often provided agricultural training to men (Waring, 1990 in Anderson, 1995: 74). These policy interventions are intended to have a positive effect on the lives of women but fail to do so because of mistaken assumptions about the nature of women's work.

3.3.2 Illustration 2: Divorce

Anderson further argues that definitions of *divorce* presuppose value judgments. She draws on the work of Stewart et al (1997), 'Separating Together: How Divorce Transforms Families', who attempt to study the concept of divorce in a way that recognises the influence of gender roles in the family.

One conceptual framework for divorce is as a separation typically with an undesirable effect (Stewart et al. 1997 in Anderson, 2004). This might guide us towards the undesirability of children living with only one parent or traumatic experiences of those experiencing divorce. This is a conception of divorce in terms of only its negative aspects (e.g., trauma and loss) (Anderson, 2004). Choosing to focus on divorce in terms of *trauma* or *loss* reflects a more traditional view of the family unit (Anderson, 1995). As we saw above, an androcentric conception of a household assumes that a household consists mainly of a male head of household and a female who is supported by the male in performing mainly childcare and other household duties (e.g., "housework") (Anderson, 1995: 72). However, Stewart et al (1997) approach divorce with greater ambivalence. For example, the researchers articulate divorce as an opportunity (e.g., for personal growth or for transforming the family) and thus reject assumptions about traditional households.

Given this value orientation, the researchers are compelled to ask individuals[16] about their experiences of divorce and how their coping strategies changed over time (Stewart et al. 1997). The researchers are more likely to consider positive aspects relevant to divorce, such as the opportunity for families to move on and adjust to new life, coping strategies over time, and so on (Anderson, 2004: 22). In doing so, the research team offers the possibility of finding evidence that different families may find alternative life plans fulfilling.

Given the background assumption that divorce is something negative, evidence more likely supports a theory that families and children are negatively impacted. Whereas alternative characterisations of divorce may not hold these background assumptions (Anderson, 2004). Given a different background assumption, perhaps one that does not accept that divorce is

---

[16] It is precisely because certain individuals have a privileged normative authority to make judgments about their own experiences of fulfilment that research programmes that draw on self-assessments and personal experiences are valuable (Anderson, 2004: 21).

always something undesirable, evidence less likely supports a theory that families and children are negatively impacted. Hence, background assumptions influence what data is afforded priority by the researcher, and it is only when they are surfaced that we realise that they influenced us to see something as more or less relevant in the process of inquiry (Anderson, 2004).

So, if you think divorce may be a good thing not an evil, you *look* for facts about whether divorce is emancipating for women and the knock-on effects on children. I suppose here that if evidence is emancipating for women and that benefits children is found, both pro and anti-divorce folks will take it to be *relevant*. On the other hand, suppose evidence that it is emancipating is found but no evidence one way or the other about effects on children. Then its *relevance* could depend on background assumptions (as Longino and others say). If some people take it as given that women having more self-fulfilment is good for their children, they will see it as relevant. Those who have no view about this will not have a reason to see this as relevant to children's welfare.

Anderson (2004) uses the example of divorce to make a distinction between *legitimate and illegitimate values* (see also Douglas, 2009; Hicks, 2014; Intemann, 2015). She claims,

> some uses of values to select background assumptions are illegitimate. Feminists object to the deployment of sexist values to select background assumptions that insulate the theoretical underpinnings of patriarchy from refutation. Critics of feminist science similarly worry that feminists will use their values in ways that insulate feminist theories from refutation. We need criteria to distinguish legitimate from illegitimate ways of deploying values in science. (Anderson, 2004: 2)

I will not go into this important literature in any detail here. But it is important to note Anderson's core proposition since it is relevant to subsequent chapters of this thesis.

Anderson claims, those who advocate for some version of value freedom do so because of "what they take to be the dogmatic character of value judgments" (Anderson, 2004: 3). This is the idea that if we allow value judgments to guide inquiry then we allow presumption and conjecture into our practices that are likely blind to evidence (Anderson, 2004). The worry is that this would impede the recognition of evidence that might undermine our preferences for a particular outcome (Anderson, 2004: 3, 9). However, for Anderson, value-laden assumptions are legitimate *if* they are amenable to empirical testing and revision, and do not predetermine the conclusion (Anderson, 2004).

For example, researchers may draw on feminist values in the collection and processing of evidence. Choosing a particular value-orientation (e.g., divorce as a transformative) may encourage researchers to look for this type of evidence but does not guarantee that they will find it (Anderson, 2004). Considering divorce as a positive evaluation is conducive to broadening the evidence base beyond only focussing on the negative aspects of divorce. Researchers seek different types of evidence, and this guides the research "to discover a broader range of evidence than would otherwise be the case" (Anderson, 2004: 20). We promote better science when we challenge normatively held values to uncover a broad range of evidence that bears on our question, as Anderson argues (Anderson, 2004 in Goldenberg, 2013: 23).

It is legitimate to conceive divorce as something positive and, on this basis, look for evidence in favour of it—so long as one does also hunt for disconfirming facts and does not discount them when they turn up. Whereas values are illegitimate when the employment of value judgments drive inquiry toward a predetermined conclusion (Anderson, 2004).
"If a hypothesis is to be tested, the research design must leave open a fair possibility that evidence will disconfirm it" rather than direct scientists towards its corroboration (Anderson, 2004: 19). When values push scientists to corroborate already endorsed claims rather than impartially assess the evidence for and against them, this threatens the reliability of belief formation in science, as it contributes to the one-sided processing of facts.

3.3.3 Illustration 3: Self-defence

Feminist standpoint theorists take Anderson and others as a starting point for the argument that some actors have a special epistemic advantage in research on particular topics. Feminist standpoint theorists claim that certain social and political positions occupied by women (and by extension other groups who lack social and economic privilege) can become sites of epistemic privilege (Intemann, 2010)[17]. Such historically disadvantaged positions offer a productive starting point for inquiry into questions of interest to feminists. For example, an individual's perspective is shaped by their social, political, and cultural experiences and so an individual’s standpoint informs which concepts are deemed appropriate, which claims are heard and understood by whom, which features of the world are salient, which reasons are understood to be relevant, which conclusions credible, and so on (Crasnow, 2014: 152-155). These ideas drive at a very particular value issue. The central idea is that different standpoints pull information together in different ways and this affects choices about whether the certain facts are ‘evidentially relevant’, or whether the evidence is sufficient to warrant a conclusion. This is taken to suggest that some actors acquire unique information about a case given their critical experience of phenomena.

Consider an illustration of feminist standpoint theory. Criminologist Elizabeth Stanko studies the various strategies that women adopt to avoid assault (Stanko, 1997 in Crasnow, 2014: 152-153). Stanko conducted her work as an “insider-outsider” (Crasnow, 2014: 152). As a woman, she was also able to anticipate how women’s behaviours affect the daily lives of women. For example, she was familiar with strategies for avoiding assault and keeping safe at night. As a sociologist, on the other hand, she was able to appreciate the way that these behaviours were shaped by the social structures in which women live (Crasnow, 2014: 153). Stanko began by asking women to offer their accounts of avoiding assault and revealed descriptions and activities that were not previously considered in traditional research, including “choosing a place to live, deciding when and where to walk, choosing a time to go to the laundromat or grocery store, deciding what to wear, and so on” (Stanko, 1997 in Crasnow, 2014: 152). Stanko’s role as an insider-outsider enabled her to put the information together in such a way that captured the relevance of this information to our understanding of how women defend themselves against assault. Crasnow notes,

> Her feminist standpoint includes awareness of how power relations shape social structures. Her results suggest the need for stretching the concept of

---

[17] Some feminists specifically want to advance feminist scientific research programmes that are explicitly informed by feminist values (Anderson, 2002). For example, Anderson argues that finding a proper role for feminist values requires “careful attention to the goals of scientific inquiry” (Anderson, 2002: 498). This is an important argument but not one that I will address here.

> self-defence in ways that lead to the identification of *a broader range of evidence* (Crasnow, 2014: 153, emphasis added).

Stanko was able to model self-defence differently than it is normally understood (Crasnow, 2014: 155). Self-defence is perhaps traditionally understood as something that occurs when women are doing something overtly dangerous, like walking down alleyways when they are alone at night (Crasnow, 2014). However, her research challenges the traditional view found in much criminology research. It is only because Stanko was aware of the value to look for certain descriptions of behaviours that are relevant to assault (e.g., late-night trips to the laundry) that they were included in the research (Crasnow, 2014).

What this example shows is that female scientists, like Stanko, are in a unique position to point to patterns of behaviour that those in a dominant group or culture are unable to recognise. Female scientists may be able to draw on their experiences as both scientists and women having experienced forms of harassment, and so occupy a privileged normative authority to make value judgments in research programmes (Intemann, 2010).

One important way that an insider-outsider contributes to knowledge is that the background assumptions of the dominant group are more transparent to the insider-outsider but are not to those who lack this experience (Intemann, 2010). The inclusion of members of marginalised groups has the potential to lead to more rigorous critical reflection because their experiences will often be precisely those that are most needed in identifying problematic background assumptions and revealing limitations with research questions, models, or methodologies (Crasnow, 2014). So, like Longino, standpoint theorists offer a role for the inclusion of different perspectives. This extends beyond the woman's perspective and includes representation from other underrepresented individuals, like non-western opinions or other often marginalised perspectives.

3.4 A new example: The Ebola Outbreak in West Africa, 2014

In the remainder of this chapter, I offer an example of my own. I will reconstruct a caricature of the Ebola outbreak in West Africa in 2014 using literature on responses to the Ebola virus pandemic in a way that fulfils an illustrative purpose only. The facts I present are informed by real empirical research conducted both at the time and after the outbreak in 2014, such as that collected by anthropologists working in the field and media reports on the incident (Wilkinson and Leach, 2016; Fairhead, 2016). However, I cannot claim I have provided a complete account. How good my example is will depend on a lot of empirical research whereas I am just reconstructing a plausible narrative that is illustrative of the real-life case. I don't know whether this is exactly what happened but rather intend to present a plausible story of how it might have gone. The reconstruction of the Ebola outbreak in West Africa in 2014 is useful for the purpose of thinking about the role of contextual values in the production of and use of evidence.

3.4.1 The collection and processing of evidence

The standard way to categorise where feminists and others see values entering is as follows: deciding what hypotheses are *significant*, deciding if a fact is evidentially relevant, deciding what to do given the underdetermination problem. Thinking about the example of the Ebola outbreak has led me to sort places for values to enter a bit differently. I consider feminist empiricist philosophers use of contextual values in what I pick out as three distinct ways:

1. the selection and collection of facts
2. the turning of facts into evidence
3. and the weaving of these together to confirm or disconfirm a hypothesis

The first category refers to the collection of facts. This refers to the process of finding out what facts exist. The second category refers to the use of facts as evidence. Nothing comes labelled as evidence for a given hypothesis. There must be a reason to see the obtaining of given fact as evidence for the truth of some other claim. I then add a third category - the weaving of these together to confirm or disconfirm a hypothesis. My third category refers to way that certain kinds of facts are emphasised in the production of evidence for a hypothesis[18].

---

[18] One way to explore such a case is using a causal-process tracing methodology. I will outline this method in more detail in chapter 5. This example is for illustration only and so it does not require the presentation of a causal-process tracing methodology.

On March 23, 2014, the World Health Organisation officially declared an Ebola outbreak after the virus spread to Guinea's capital, Conakry. The outbreak was the largest in history and spread from Guinea to nearby Sierra Leone and Liberia (World Health Organisation, 2021). The World Health Organisation Director-General Margaret Chan framed the frightening situation as the "most severe acute public health emergency in modern times" (Wilkinson and Leach, 2015: 136). What is interesting about the West African Ebola outbreak is that it revealed failings in the ability of the World Health Organisation to respond to outbreaks. A consequence of this has been the expansion of approaches aimed at strengthening international health surveillance and security. These are approaches that seek to reduce vulnerability to acute public health events, commonly infectious disease outbreaks like Ebola, that affect populations across international settings. The approaches put forward by the World Health Organisation do not only focus on the primarily low and lower-middle income countries in which the World Health Organisation are present but focus on protecting other countries (including the United States and the United Kingdom) from infectious viruses like Ebola. As such, health surveillance approaches can be ethically, culturally, and politically challenging.

Countries in equatorial Africa have experienced Ebola outbreaks since 1976[19]. All previous outbreaks, which remained largely confined to remote rural areas, were controlled, with support from the World Health Organisation and other international partners, in periods ranging from three weeks to three months. According to the World Health Organisation, clinicians in equatorial Africa have good reasons to suspect Ebola when a "mysterious" disease occurs, and this favours early detection. Laboratory capacity is in place. In past outbreaks, amplification of infections in health care facilities was the principal cause of initial explosive spread. Staff know where to send patient samples for rapid and reliable diagnosis. Health systems are familiar with Ebola and much better prepared. For example, hospitals in Kinshasa, the capital of the Democratic Republic of Congo, have isolation wards, and staff

---

[19] The Ebola virus is deadly. Its origins trace back to 1976 near the Ebola River in the now Democratic Republic of Congo. The initial case of the most recent Ebola outbreak (or index patient) was reported in December 2013 in a small village in Guinea (CDC, 2019).

are trained in procedures for infection prevention and control. Governments know the importance of treating a confirmed Ebola case as a national emergency.

In contrast to equatorial Africa, West African countries have not previously experienced an Ebola outbreak prior to 2014 and so were poorly prepared. According to the World Health Organisation and other international partners, clinicians had never managed cases. No laboratory had ever diagnosed a patient specimen. No government had ever witnessed the social and economic upheaval that can accompany an outbreak of this disease. In the early stages of the pandemic, few health professionals were aware of the spread of Ebola, and it was mistaken for other infectious diseases, such as malaria, typhoid fever, yellow fever and hemorrhagic fever. Limited access to screening and diagnosis, poor pre-existing surveillance, overstretched isolation facilities and a lack of treatment have been cited (Rugarabaum et al. 2020). Poor public health infrastructure in West Africa was largely understood to influence the rapid spread and transmission of the deadly Ebola virus across communities in a short timescale. Furthermore, because of the nature of transmission[20], the Ebola virus was seen as particularly susceptible to sociocultural events such as caring for the sick, burials, funerals, and mourning practices. There was huge emphasis put on the importance of what was described as 'safe and dignified burials' which was also understood to play a crucial role in preventing future virus transmission (World Health Organisation, 2014).

In 2016 it was recognised that the "the [official 2014 Ebola outbreak] response suffered from insufficient local technical capacity and deficient understanding by international partners of the fundamental needs and culture of the affected communities" (World Health Organisation, 2016). One criticism levelled against the international effort to handle the Ebola outbreak concerns the language used to frame the Ebola response. The World Health Organisation in particular was criticised for focusing on 'poor health systems' and 'weak surveillance structures' at the expense of other explanations, such as more social or historically defined narratives (Wilkinson and Leach, 2014). The problem was not that the

---

[20] The virus spreads to people initially through direct contact with the blood, body fluids and tissues of animals. Ebola virus then spreads to other people through direct contact with body fluids of a person who is sick with or has died from Ebola Virus Disease (EVD). This can occur when a person touches these infected body fluids (or objects that are contaminated with them), and the virus gets in through broken skin or mucous membranes in the eyes, nose, or mouth.

statements on offer were untrue, but that they offered an incomplete account of the context of Ebola. It is worth going back to reconsider the ways in which the official response reflected the 2014 West African Ebola outbreak.

At the time of the outbreak, the World Health Organisation adopted this hypothesis about its spread: the spread was heavily due to the fact that local health systems in Guinea could not cope with the Ebola virus (World Health Organisation, 2021). This was based on two main facts. Fact 1: the susceptibility of the virus to weak health infrastructure and traditional practices, such as caring for the sick, burials, funerals, and mourning practices (World Health Organisation, 2021). For example, the World Health Organisation found studies to confirm infections could be attributable directly to participation in care or mortuary rites (World Health Organisation, 2014). Fact 2: intensification of transmission was found to be a result of few health professionals being aware of the spread of Ebola, together with the virus being mistaken for other infectious diseases, such as malaria, typhoid fever, yellow fever and hemorrhagic fever (World Health Organisation, 2014).

Anderson argues that the extensions of the phenomena that we are interested in is defined by our own social and political values (Anderson, 2002: 506). This is true with my example of the 2014 West African Ebola outbreak. What I want to suggest is how the kind of conceptualisation that the World Health Organisation brought with it affected the hypothesis it leaned to and then what would count as evidence for that. The World Health Organisation, for example, is responsible for coordinating international health in outbreaks, and so it was natural to focus on Ebola in terms of health systems, preparedness, surveillance, security, and response (Wilkinson and Leach, 2015). The World Health Organisation put forward a body of evidence in support of the role of poor public health infrastructure in explaining the Ebola outbreak (World Health Organisation, 2015). This included lack of previous experience with Ebola, limited access to screening and diagnosis, poor pre-existing surveillance, overstretched isolation facilities and a lack of treatment options, and so on (World Health Organisation, 2014).

One potential problem was the focus on local health systems not being able to cope with the Ebola virus. What I want to suggest is that this may have led to the misrepresentation or dismissal of facts that for someone else, with different starting assumptions, might count as evidence. According to Fairhead (2016), there was a lack of attention to the cultural aspects of the epidemic (Fairhead, 2016). Different characteristics focus on distinctive aspects of the Ebola outbreak, including personal accounts of community resilience in healthcare (Wilkinson and Leach, 2015). Also important are cultural characteristics, which include the role of community health leaders or traditional and/or spiritual leaders in the spread of the Ebola outbreak. According to Wilkinson and Leach, for instance, the focus on *weak* health security under emphasised the material conditions of poverty and history of violence that underpin community behaviour. Despite facts emerging that sick people were staying in communities, there was an official resistance by the World Health Organisation to exploring the factors surrounding social resistance to Ebola response (Wilkinson and Leach, 2015). As anthropologist Anoko pointed out at the time of the outbreak, “the problem was that the people handling the intervention only looked at this as a health issue; they did not try to understand the cultural aspects of the epidemic” (Anoko, 2014 in Fairhead, 2016). Following Fairhead,

> The public health logic was clear: the ill should be isolated within Ebola Treatment Centres (ETCs) and burials should be made safe. Yet this message was not heeded by all, and for many reasons. Symptoms of Ebola were often hard to distinguish from many other ailments, and people delayed traveling to ETCs until it was too late. News that Ebola was deadly and had no cure did not make Ebola Treatment Centres attractive, and travel was difficult to arrange and costly in any case. Families worried that children and others would not survive the quarantine that would be imposed on any family in which the virus had been identified. And misinformation that Ebola was caught from bushmeat provided false confidence to those who didn’t hunt or eat it (Fairhead, 2016: 8).

One assumption adopted by the World Health Organisation at the time of the outbreak was that the focus on health surveillance systems is a good explanation for understanding the Ebola outbreak. This presupposes the value of international surveillance systems in explaining the outbreak of Ebola in communities. Explaining the outbreak as a health surveillance problem informs what facts are gathered and counted as evidence in favour of a hypothesis, which is in turn affected by a dominant perspective and the values encoded in that perspective. I will argue that certain facts may be overemphasised in the production of evidence for a hypothesis. Accepting the hypothesis too soon is a failure to investigate the facts. Finding out what facts exist cannot be separated from the interests of the inquirer (Anderson, 2002). This is one important way that values enter in the production of evidence for a hypothesis.

Consider these more detailed hypotheses (or claims) that the World Health Organisation accepted[21]. Hypothesis 1: conducting 'safe and dignified burials'[22] will reduce the spread of the Ebola virus. The evidence for the hypothesis is based on two facts. Fact 1: Infection of the Ebola virus was found to occur from the touching of bodies of those who have died from the disease and so to prevent the spread (World Health Organisation, 2017). Fact 2: The handling of dead bodies was also found to be a major contributing factor in the outbreak (World Health Organisation, 2017). The problem here is familiar - that the hypothesis is underdetermined by the evidence in that the evidence available to us is insufficient to determine what beliefs to hold in response to it. If we take it for granted that ideals about the handing of the dead are shared across different cultures then, given those background beliefs, we are likely to take these two facts as evidence in support of the hypothesis. If, on the other hand, we think that ideals about death are *not* widespread, or if we think it likely that some cultures will reject the World Health Organisation notion of a 'dignified' burial, then, given *those* background beliefs, we may seek additional facts. The most obvious next step is to gather more evidence by investigating the facts.

Facts concerning spiritual practices were not collected because of taken for granted assumptions about the acceptability of burial practices. Two practices ensured that connections between the medical officers and local communities were likely to break down (Fairhead, 2016). One was the fact that bodies were returned unwashed and in zipped-up body bags to be buried by the Guinean Red Cross (Fairhead, 2016). Another was the fact

---

[21] Note that the hypotheses that I suggest are just exemplary of the types of claims that the World Health Organisation adopted at the time. They are not official claims made by the World Health Organisation, although the facts used to support my hypotheses are based on actual data.

[22] A safe and dignified burial procedure included several steps: the placement of a body in a body bag, placement of a body bag in a coffin where culturally appropriate, sanitizing of the families' environment, burial at the cemetery engaging community for prayers to dissipate tensions and provision of respectful time, among other steps (World Health Organisation, 2017). The handling of dead bodies was advised to be kept to a minimum and burials were recommended to be undertaken by the safe burial teams (World Health Organisation, 2017).

that bodies were not being released to the care of their families (Fairhead, 2016). According to the African belief system, life does not end with death, but continues into another realm. A well-cited example is that of a pregnant Guinean woman whom mourners refused to allow a team of outsiders to bury. If she was to be saved from eternal wandering and reach the village of the dead, it was insisted that her foetus must be removed. In traditional Kissi culture, a woman buried without her foetus disturbs the world's natural cycles which begin and end. The failure to remove a foetus poses a huge risk to an individual or community since it risks disrupting natural cycles which have a beginning and an end (Fairhead, 2016).

It was taken for granted that everyone's priorities were the same. This is an assumption that everyone sees the problem the same way. The World Health Organisation prioritises containment, safe disposal, and international health security. Whereas communities clearly saw the problem differently. Death from a virus like Ebola is not the most pressing concern for many people in affected areas; fear of sexual violence, malaria and other diseases are often more immediate concerns (Wright and Parker, 2020). Moreover, people want to make sure that their family members and communities die properly (Wilkinson and Leach, 2016).

Given our background assumptions we find different evidence relevant to the hypothesis under scrutiny. From the perspective of international health security, we emphasise evidence on weak infrastructure and health knowledge (Wilkinson and Leach, 2015). However, with a commitment to alternative social or cultural perspectives, we emphasise different evidence. Evidence on the specific roles played by communities in the handing of the dead, and the challenges they faced, were missed. So, one place to start would be to investigate whether or not the affected communities' express agreement with burial practices. Additional facts might include religious burial practices including the washing of dead bodies, the spiritual significance of hugging the bodies of elders in preparation for burial, and so on (Wilkinson and Leach, 2015: 139). For example, we may find evidence of widespread disapproval of burial procedures that do not respect the afterlife and spiritual beliefs. For example, Leach argues that international actors like the World Health Organisation and its partners knew that funerals were a key transmission route for Ebola with family and communities coming from long distances to take part in burial rituals, but they were making little headway in preventing this practice from occurring. Rather than working with the dominant approach put forward by the World Health Organisation, communities were actually finding their interventions offensive.

By recognising the social significance of burial attendance, we seek facts specifically on community trust with regards to burials. Including facts on the local leaders involved in burials, substituting physical and non-physical rituals, the issue of gaining agreement and establishing acceptance and trust within the community to delay traditional visits until after the crisis. Or facts to exemplify the importance of touch and the senses in the dealing with the death of loved ones within a community. Considerable emphasis is put on the continuation of life into death, with a focus on natural beginnings and endings as is characteristic of African belief systems. For example,

> Most people hope that at death they will be born back into a parallel “village of the dead” where they will be reunited with those who had predeceased them—and also that they will retain links with the living. Living relatives conduct rites to encourage this result, because there is nothing more problematic for the living than an aggrieved ancestor, or one who has been unable to pass to this other world. Cherished relatives are buried in the village, and even a relative who is buried elsewhere can be brought home in the form of an object (often an antique iron bar) that was placed in and then removed from the actual tomb. Care is needed, however, to ensure that those who might bring ill fortune to the living are buried elsewhere and not returned to the village in any way. Sorcerers cannot be sent back to this parallel village, where they might continue their evil. Suspicions that a dead person was a sorcerer need to be investigated prior to their burial, and there are modes of autopsy that ascertain this. Nor is it acceptable to reintroduce into this village of the dead those whose “bad” death (for example by lightning strike) call their good character into question, or to do so for strangers whose character is unknown (Fairhead, 2016: 14).

There are different understandings of the causes of viruses, where a traditional doctor will look for the cause of the virus as misfortune between the patient and the social and spiritual environment. As Wilkinson and Leach (2015) note, these are matters of life and death and social control over life and death. West Africa, and especially its rural communities, is a region where caring for the sick and ensuring that funerals are properly conducted is critical to enabling people when they die to become proper ancestors. It is also crucially important to settling affairs around property and inheritance. Properly understanding the causes of death is crucial for communities in West Africa so that people are not accused of being sorcerers and causing death through nefarious means.

Facts concerning spiritual practices and facts concerning the specific roles played by communities were not woven together properly to confirm or disconfirm the World Health Organisation's central hypothesis that 'conducting 'safe and dignified burials' will reduce the spread of the Ebola virus'. This is because of the taken for granted assumption that ideals about the handing of the dead are shared across different cultures[23] (World Health Organisation, 2017). We may *under emphasise* certain facts in our assessments, such as the role of spirituality, if we uncritically accept the assumption that beliefs about death are shared across cultural groups. Facts may appear compelling, but there are different ways to explain something and facts relevant to these different conceptions also need to be collected. It is good scientific practice to rule out alternatives.

It is important to note that there are different perspectives within the West African context. In a given setting, not everyone will accept "traditional" belief systems, and there will be a range of perspectives. There are key differences in burial practices between groups within the West African context. The purpose of burial rituals is threefold: to honour the dead relatives in a traditional way; to say good-bye to loved ones to accommodate the deeply held beliefs and obligations of the living to the dead, respecting the cultural view of life after death. One of the main rituals that seems to be common across all West African groups is the washing of the body. However, certain additional burial traditions depend on the tribe and religion. For example, there are special circumstances, such as when cultural groups feel that the foetus needs to be removed from the mother's body before burial, as we see above. If these procedures are not undertaken properly, there are consequences for both the deceased and the living relatives, as per the cultural beliefs of some groups.

Nevertheless, there may be cases where there is a clash between local beliefs and the health policy implemented by the WHO. In such cases it is important not to pretend that the beliefs

---

[23] The World Health Organisation explicitly refers to burials as culturally sensitive. The WHO does not explicitly commit to addressing and overcoming differences in fundamental belief systems; at least at the time of this thesis.

found in many traditional African systems are compatible with contemporary scientific worldview. The implicit view in my argument is that by better communication it would be possible to have the WHOs safe burial practices implemented. However, there are cases where there will be a direct clash between a traditional practice and an unavoidable health necessity. For example, health authorities concerned with overall health may choose to err on the side of caution and ensure that all corpses are buried in body bags with sterile procedures – regardless of beliefs of traditional burial requirements. On the other hand, local people may not value health outcomes above the way they wish to honour their dead. Sometimes it may be possible to negotiate a middle ground and honour reasonable requests, but there can well be irresolvable value conflicts. I am not offering a solution to that problem.

Next consider hypothesis 2: the establishment of Ebola Treatment Units and the targeted strengthening of weak health and surveillance systems, including investment and training of health staff and surveillance technology will reduce Ebola virus transmission. The evidence for the hypothesis is based on two facts. Fact 1: investment in surveillance technology provides opportunity for better insight as to where cases of Ebola exist (World Health Organisation, 2015). Fact 2: the number of established facilities and placement of medical staff in Ebola Treatment Units aids response efforts (World Health Organisation, 2015).

Evidence against the hypothesis would be based on further facts. For example, facts concerning a lack of community acceptance and mistrust in the medical health infrastructure as advocated by the World Health Organisation and other intervening health agencies.

If we take it for granted that community acceptance of health surveillance technology is both widespread and of the sort that leads to effective intervention, then, given those background beliefs, the emphasis we put on facts 1 (e.g., investment in surveillance technology provides opportunity for better insight as to where cases of Ebola exist) and 2 (e.g., the number of established facilities and placement of medical staff in Ebola Treatment Units aids response efforts) will be quite high. If, on the other hand, we hold the belief that the community acceptance of health surveillance technology is *not* widespread, or if we think it likely that communities may avoid reporting instances of Ebola and/or treatment, then, given *those* background beliefs, the emphasis we put on finding additional facts (e.g., mistrust in medical infrastructure) will be quite high. The most obvious next step is to gather more evidence by investigating the facts. An obvious place to start is to investigate community resistance to health surveillance and/or what other networks of support are in place.

At the time of the 2014 outbreak, one important source of information was the cultural perspective of the community health workers involved in the containment effort in West Africa in 2014. Many community health workers remained active in their communities and played important roles in the Ebola response despite a lack of formal engagement or support from internationally supported health efforts. The specific roles played by community health workers, and the challenges they faced, are however not well documented. In fact, the accounts of community health workers and local volunteers are barely visible. Yet it is these accounts that provide valuable insights into the social, psychological, and cultural effects of Ebola that go against the dominant political and humanitarian rhetoric (Kingori and McGowan, 2016). Facts related to community systems of care characteristic of many West African communities may not be collected if we take for granted the importance of a health infrastructure supported by community health workers and local volunteers. Here, I am referring to a whole body of information that may be missed.

Further consider the history of negative local interactions with external actors, especially interventions imposed by international health agencies as Fairhead does. Fairhead (2016) describes that any attempt to push forward external messages that didn't take on board the existence social and cultural practices were met with hesitation and mistrust towards outsiders:

> anthropological literature on the [West African] region has revealed varied speculations and suspicions about the nefarious ways in which Euro- Americans (and others from around the world) acquire their wealth: how they have occult methods of identifying diamondiferous and auriferous regions; how they have "four eyes" just as indigenous "seers" do; how they have devices (sometimes described as binoculars) that help them see beyond the surface of things; how they acquire African masks to capture and harness their spirits; how they catch

> nature spirits, transporting them to Europe for profit, just as they have been exporting diamonds and other precious resources for decades (e.g., Diggins 2014; Diawara 1994). In short, although rarely stated, the people of the Forest Region have long coexisted with these “antisocial” intruders, but have reached an accommodation of sorts with those wielding such powers (Fairhead, 2016: 22)

Sierra Leone’s health system, for example, reflects a similar historic struggle (Wilkinson and Leach, 2015). Sierra Leone historically consists mostly of farming villages and distances to the health clinics are vast and require private transport or hammocks to carry the sick (Wilkinson and Leach, 2015: 139). In the rainy season this is more difficult, and paths become impassable. Informal healthcare also forms much of the infrastructure and the legitimacy of these providers is often greater than of the healthcare workers provided by the government (Wilkinson and Leach, 2015: 139). Communities put great trust in elders and local providers of care. Moreover, fear, misunderstanding and stigmatisation are all factors that contribute to community mistrust of healthcare systems, and poor health seeking behaviour. The assumption that established medical practices will be accepted in communities in West Africa risks overlooking what’s there already in terms of community apparatus and historical know-how. Especially considering that the wider context in which the Ebola outbreak occurred is characterised by diverse religious and spiritual values.

Facts 1 (e.g., investment in surveillance technology provides opportunity for better insight as to where cases of Ebola exist) and 2 (e.g., the number of established facilities and placement of medical staff in Ebola Treatment Units aids response efforts) may appear compelling if we take it for granted the community acceptance of western style healthcare. But there are different ways that the story could have gone. The drawing together of facts from different explanations tells a different story. Facts 1 and 2 were not considered against additional facts (e.g., on community mistrust). So, the facts were not woven together properly because of taken for granted assumptions about community acceptance in health systems. We may overemphasise the health systems explanation over the alternative explanations if we remain committed to biasing assumptions.

Finally, consider hypothesis 3: community resistance[24] is due to lack of awareness on the risks of exposure to the virus. To offer some context, resistance began in Guinea, Sierra Leone, and Liberia (the three worst affected states) as medical officers tried to separate the dying and the dead from their communities (World Health Organisation, 2015b). There were widespread reports of ambulances and burial teams being bombarded with stones (Maxmen, 2015). For example, “within days of Ebola being identified in March 2014, MSF had opened an ETC in Macenta town. Yet within a week it was attacked by an angry crowd” (Fairhead, 2016: 13). Other attacks were more vicious still, including the deaths of medical officers (Maxmen, 2015). Violent resistance was initially thought to be due to a lack of education on the cause of the virus and other cultural misunderstandings (Fairhead, 2016). The evidence for the hypothesis was based on two key facts. Fact 1: violent resistance at the community level by those without formal education (World Health Organisation, 2015b). Fact 2: widespread misinformation on the causes of the virus and its links with the consumption of bushmeat (Muller, 2021).

In order to take facts 1 (e.g., violent resistance at the community level by those without formal education) and 2 (e.g., widespread misinformation on the causes of the virus and its links with the consumption of bushmeat) as evidence of the hypothesis, we have to assume that education is a valuable proxy for awareness of risks. If, on the other hand, we see the instances of conflict as a reason to step back and ask what assumptions are misplaced, then, given *those* background beliefs, the emphasis we put on this fact informing the hypothesis will be low. Our assumptions define what kind of evidence is understood to be relevant. If we reject the assumption that education is a valuable proxy for understanding risk, then the most obvious next step is to gather more evidence by investigating the facts. Given an alternative perspective, one that sees community resistance as a conflict with outside influence in community-oriented health systems, we find additional evidence relevant to the hypothesis. An obvious place to start is to investigate community resistance to primary healthcare and international intervention. Equally important is the history of negative local interactions with international (outside) interventions in local community

---

[24] An additional point not addressed here is the use of biasing terminology: consider terms such as ‘community resistance’ and ‘population resistance’ that privilege the perspective of the World Health Organisation over the more nuanced concerns of the communities.

settings, especially those imposed by international health agencies. So, we need the following:

Fact 1: violent resistance at the community level by those without formal education.
Fact 2: widespread misinformation on the causes of the virus and its links with the consumption of bushmeat.
Fact 3: community resistance to international (outside) intervention
Fact 4: the history of negative local interactions with international intervention, especially those imposed by international health agencies.

Facts 1, 2, 3 and 4 are not woven together properly. Facts 1 and 2 may appear compelling if we take it for granted the community acceptance of western style healthcare. The use of facts 1 and 2 in isolation fails to understand the significance of religious and multicultural issues in community resistance. We may under emphasise facts 3 and 4 in our assessment of the hypothesis if we take for granted the community acceptance of western style healthcare. Additional information may be necessary to confirm the hypothesis. Cultural and religious knowledge can provide the sort of insight that forces critical scrutiny and makes it more likely that background assumptions and perspectives will be justified. We want to adequately test a hypothesis from different explanations, we don't want to favour a preferred perspective. Facts may appear compelling, but there are different ways that the story could have gone and facts relevant to this also need to be collected. It is good scientific practice to rule out alternatives. Given that hypotheses are always underdetermined, claiming that the evidence supports the hypothesis without further consideration of at least plausible alternatives from different perspectives is a failure to investigate the facts.

As feminist philosophers argue, scientists express the contextual values inherent in their cultures and particular lives when they engage in a scientific activity (Longino, 2002: 135-136). These include beliefs about what is acceptable or unacceptable practice, which may vary from community to community (Longino, 1990). One way to improve matters can be to try to ensure the values inherent in our assumptions are made explicit so that we better understand the kinds of roles that values are playing in scientific reasoning.

According to feminist philosophers, there are assumptions that can be moral or cultural which may seem natural from one moral point of view from another. For example, the assumption that what we think of as a safe and dignified burial is shared across cultures. Assumptions are influenced by a particular worldview and yet these enter in the selection of facts and in their use as evidence in support of a hypothesis. In addition, assumptions may reflect a cultural bias that prevents us from seeing facts as relevant. We need to be cognisant of the background assumptions on which our judgments depend. We can guard against inheriting unjustified assumptions by being conscious of the limitations of our own perspectives. This requires those leading an international response to find ways to critically assess the most basic assumptions underpinning a particular perspective. One help offered by feminist philosophers is open discussion to try to uncover the assumptions that drive ours and others' thinking. It's only by making these assumptions explicit that we have a chance of meaningful dialogue. This is the kind of discussion that gets leading institutions to change their minds in light of evidence and acknowledge when the facts lie out of reach and accept the remaining uncertainty.

What I have highlighted in my example is that there were shortcomings in the official response to the Ebola outbreak in 2014. My example outlines the need to better understand the concerns of affected populations in an official response to an outbreak. This includes understanding the needs, priorities, and concerns of those most affected. The value of community knowledge needs to be recognised and included from the very beginning. This is a concern about how best to integrate different value schemes in the production of evidence. Or how best to integrate different value schemes into a response to an outbreak.

3.5 Conclusion

I defend a very particular role for values in the collection and processing of evidence: in the selection and collection of facts, the rendering of facts as evidence and the weaving of these together to confirm or disconfirm a hypothesis. Background assumptions pull facts together in different ways and this affects whether the certain facts are evidentially relevant. What facts we take to be evidence for (or against) a hypothesis is mediated by background assumptions, which is one vehicle through which contextual values enter. Contextual value judgments play a legitimate and ineliminable role in the decision to accept one hypothesis over another in that they help fix what facts are relevant as evidence—though they should not fix whether (once taken as evidence) they confirm or disconfirm the hypothesis. Values are illegitimate when they push scientists to confirm already endorsed claims rather than impartially assess the evidence for and against them (Anderson, 2004; Douglas, 2016). This leads to the one-sided processing of facts.

We need to make sure that the background assumptions brought to inquiry do not determine the hypothesis in advance but leave it open to determination by the evidence (Anderson, 2004: 11). Facts may appear compelling, but there are different ways that the story could have gone and facts relevant to this also need to be collected. Good scientific practice requires us to rule out alternatives.

References

Anderson, E. (1995) Feminist Epistemology: An Interpretation and a Defence, Hypatia, 50-84.

Anderson, E. (2002) Situated Knowledge in the Interplay of Value Judgments in Scientific inquiry, in P. Gardenfors, J. Wolenski and K. Kijania-Plackek, eds., In the Scope of Logic, Methodology and Philosophy of Science, vol. 2, 497-517.

Anderson, E. (2004) Use of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce, Hypatia, 19, (1).

Anoko, J. (2014). "Communication with Rebellious Communities during an Outbreak of Ebola Virus Disease in Guinea: An Anthropological Approach." Available at: http://www.ebola-anthropology.net (Accessed 29th June 2021).

Atkinson, A. B. (2001) The Strange Disappearance of Welfare Economics, Kyklos, 54, (2/3), 193-206.

Biddle, J. (2013) State of the field: Transient underdetermination and values in science, Studies in History and Philosophy of Science, 44, 124-133.

Chopra, D. (2014) Towards Gender Equality with Care-Sensitive Social Protection, IDS Policy Briefing 49, Brighton: IDS.

Crasnow, S. (2014) Feminist Standpoint Theory, Philosophy of Social Science: A New Introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford.

Deaton, A. and Cartwright, N. (2018) Understanding and misunderstanding randomised controlled trials, Social Science and Medicine, 210, 2-21.

Di Bucchianico, Maria Elena (2009) *Modelling high temperature superconductivity: A philosophical inquiry in theory, experiment and dissent.* PhD thesis, London School of Economics and Political Science (United Kingdom)

Duhem, P. (1956) Aim and Structure of Physical Theories, (first published 1914) trs. Phil- ip Weiner. Princeton. N: Princeton University Press.

Fairhead, J. (2016) Understanding Social Resistance to the Ebola Response in the Forest Region of the Republic of Guinea: An Anthropological Perspective, African Studies, Review, 59, (3).

Intemann, K. (2010) Twenty-Five Years of Feminist Empiricism and Standpoint Theory: Where are we Now?, Hypatia: A Journal of Feminist Philosophy, 25, (4): 778-796.

Intemann, K. (2005) Feminism, underdetermination, and values in science, Philosophy of Science, 72: 1001-12.

Kabeer, N. (2012) Women's Economic Empowerment and Inclusive Growth: Labour Markets and Enterprise Development, SIG Working Paper, 2012/1, International Development Research Centre/Development for International Development.

Khosrowi, D. (2019) Trade-oofs between epistemic and moral values in evidence-based policy, Economics and Philosophy, 35, (1), 49-78.

Kingori and McGowan (2016) After the End of Ebola. Available at: http://somatosphere.net/2016/after-the-end-of-ebola.html/ (Accessed 29th June 2021)

Longino, H. (1979) Evidence and hypothesis: An analysis of evidential relations, Philosophy of Science, 46, (1), 35-56.

Longino, H. (1990) Science as Social Knowledge. Princeton, NJ: Princeton University Press.

Longino, H. (1992) Taking Gender Seriously in Philosophy of Science, PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, (2), 333-340: The University of Chicago Press on behalf of the Philosophy of Science Association.

Longino, H. (2002) The Fate of Knowledge. Princeton, NJ: Princeton University Press.

Longino, H. (2014) Individuals or Populations? Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 102-123.

Maxmen, A. (2015) How the Fight Against Ebola Tested a Culture's Traditions, National Geographic. Available at: https://www.nationalgeographic.com/adventure/article/150130-ebola-virus-outbreak-epidemic-sierra-leone-funerals-1 (Accessed 29th June 2021)

Muller, S. M. (2021) Carnistic Colonialism: A Rhetorical Dissection of "Bushmeat" in the 2014 Ebola Outbreak. Available at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.656431/full#B69 (Accessed 29th June 2021).

Quine, W. V. O. (1951) "Two Dogmas of Empiricism", Reprinted in From a Logical Point of View, 2nd Ed., Cambridge, MA: Harvard University Press, pp. 20–46.

Sandmo, A. (2017) Tony Atkinson 1944-2017: A lifetime commitment to the study of inequality, The European Journal of the History of Economic Thought, 24: 3, 612-623.

Stanko, E. A. (1997) Safety Talk: Conceptualising Women's Risk as a "Technology of the Soul", Theoretical Criminology, 1: 479-499.

Stewart, A. J., Copeland, A. P., Chester, N. L., Malley, J. E., & Barenbaum, N. B. (1997). Separating together: How divorce transforms families. Guilford Press.

Waring, Marilyn. 1990. If women counted. San Francisco: HarperCollins.

Wilkinson and Leach (2015) Briefing: Ebola Myths, Realities, and Structural Violence, African Affairs, 114, (454), 136-148.

World Health Organisation (2014) Ebola Response Roadmap Situation Report. Available at: http://apps.who.int/iris/bitstream/handle/10665/137424/roadmapsitrep_31Oct2

014_eng.pdf;jsessionid=954615B086633BFBD94675D4433374D4?sequence=1 (Accessed April 1, 2021).

World Health Organisation (2015) Surveillance strategy during Phase 3 of the Ebola response, Emergency Guidance. Available at: http://apps.who.int/iris/bitstream/handle/10665/192997/the World Health Organisation_EVD_Guidance_Sur_15.1_eng.pdf?sequence=1 (Accessed April 1, 2021).

World Health Organisation (2015b) Factors that contributed to undetected spread of the Ebola virus and impeded rapid containment. Available at: https://www.who.int/news-room/spotlight/one-year-into-the-ebola-epidemic/factors-that-contributed-to-undetected-spread-of-the-ebola-virus-and-impeded-rapid-containment (Accessed April 1, 2021).

World Health Organisation (2016) An R&D Blueprint for action to prevent epidemics: Plan of Action, May 2016. Available at: https://reliefweb.int/report/world/rd-blueprint-action-prevent-epidemics-funding-coordination-models-preparedness-and (Accessed June 28, 2021).

World Health Organisation (2017) How to conduct safe and dignified burial of a patient who has died from suspected or confirmed Ebola or Marburg virus disease. Available at: https://www.who.int/publications/i/item/the World Health Organisation-EVD-Guidance-Burials-14.2 (Accessed April 1, 2021).

World Health Organization (2021). Ebola Virus Disease. Available at: https://www.who.int/health-topics/ebola#tab=tab_1 (Accessed April 1, 2021).

Chapter 4: The role of value judgment in measurement

4.1 Introduction

The goal of measurement is to get reliable data. The general public and politicians regularly expect measurement to be objective in the sense that the data is not influenced by value judgment. Value judgments reflect moral, political and cultural values and other points of view and often include what we would normally judge morally relevant norms and preferences (Longino, 1990: 216). Objective measurement can be a matter of following the right procedures, something that Daston and Galison (2007) call ‘procedural objectivity’. This is when a process allows for one scientist to be replaced with another without altering the data that is produced. With procedural objectivity, scientists rely on methodological procedures to prevent the risk of individual bias, personal interests, or value judgment influencing measurement. While it may be assumed that an objective measurement tells us something reliable about a concept that is of interest to us, I will put forward a variety of examples from across the social sciences to demonstrate that the measurement of concepts for policy is often intrinsically influenced by value judgments.

As with previous chapters, I will use a particular definition of values. This is what Helen Longino calls “contextual values” that can play in the process of scientific legitimation of claims, which many have urged should be value-free (Longino, 1990: 216). These may, and often do, include what we would normally judge morally relevant norms, preferences, like the preference to serve one social group over others.

4.1.1 Chapter outline

There are many ways of carving up the literature on measurement in the social sciences. For the purpose of this chapter, the focus is on places where values influence how we characterise, represent and formulate the procedures for a concept that we are interested in within a measure (Bradburn, Cartwright and Fuller, 2016: 3). This is of course an oversimplification of how social science really works, but it is helpful to use concrete examples for discussion. First, I consider the measurement of *civil war* as outlined in the work of Nancy Cartwright and Rosa Runhardt (2016). Second, I consider the measurement of *inflation* as outlined by Julian Reiss (2008, 2013). What is interesting about these examples is that there is no unequivocal way to display the phenomenon we are interested in when we talk about civil war or inflation. Rather the extensions of the phenomenon we offer reflects our purpose in dealing with the concept. Choosing a purpose can seldom be value-neutral. Given the uses to which the measure will certainly be put, there will be some harms and some benefits. As we will see, our choices have a knock-on effect of some issues being better represented than others (Cartwright and Runhardt, 2014). I use these examples specifically because they span the social sciences and offer a useful introduction to the topic of measurement and values, which I build on in section two of this chapter. These are relevant and important examples cited in many introductions to the philosophy of social science and values.

My contribution to this topic is the consideration of a new example of *the Hunger and Nutrition Commitment Index* developed by researchers at the Institute of Development Studies (te Lintelo et al. 2013, te Lintelo et al. 2020; te Lintelo, 2019). This is an international comparison tool, which ranks governments on the level of political commitment to address hunger and undernutrition. The index is intended for use by nutrition stakeholders (te Lintelo et al. 2013). What is particularly interesting about the Hunger and Nutrition Commitment Index is that it was developed using a co-production approach, where nutrition stakeholders were involved alongside social scientists in the selection of indicators and the weighting of those indicators. I outline four choices in the design of the HANCI and demonstrate that they are value-laden. My goal in doing so is to better understand the role of values in the production of data for use in nutrition policy.

I have three intentions in this chapter:

- To offer diverse examples from the social science literature (including civil war and inflation) that clearly demonstrate a role for values in the measurement of different social science concepts.
- To outline the kind of role that values can play at different stages of measurement (e.g., how we characterise, represent, and formulate the procedures for a concept) and what the implications are.

- To focus specifically on the production of a ranking tool to measure the political commitment to address undernutrition in different countries where values are necessary (and unavoidable) to settle the design of a measure.

4.2 The Representational Theory of Measurement (RTM)

As a way of analysing different measures within the social sciences, I apply the so-called 'representational theory of measurement' (RTM) that has become standard in philosophy of science following the significant work by philosopher Patrick Suppes and social scientist Duncan Luce, psychologist S.S. Stevens, and other philosophers of science[25]. There is a vast literature explaining the RTM, illustrating it, and using it to reconstruct specific measures in both natural and social science, including three thick volumes by Suppes and Luce themselves. This a particular theory of measurement that is standard in philosophy of science and offers a very good covering theory[26].

Suppes calls for representation theorems because we are supposed to be able to show that the measurement procedures and their formal representations are appropriate to inform us about whether measured items do or do not fall under the concept that is to be measured, or to what degree they do so. In other words, the procedures and formal representation connect with the concept. In the section below, I make use of a recent succinct account of RTM from philosopher of science Nancy Cartwright and social psychologist Norman Bradburn, who offer a three-component approach to measurement – characterisation, representation, and procedures (Bradburn, Cartwright and Fuller, 2016).

1. We define the concept, identifying its boundaries, fixing which features belong and which do not – '*characterisation'* (Bradburn, Cartwright and Fuller, 2016: 3).
2. We define a metrical system that appropriately represents the quantity or concept – '*representation'* (Bradburn, Cartwright and Fuller, 2016: 3).

---

[25] Thank you to Nancy Cartwright and Travis Chamberlain for sharing work with me on Representational Theory of Measurement.

[26] Cartwright (2008)

3. We formulate rules for applying the metrical system to tokens to produce the measurement results – '*procedures'* (Bradburn, Cartwright and Fuller, 2016: 3).

In a good measure, all three ingredients must mesh together coherently such that the characterisation is appropriate to the representation and the operational procedures are appropriate to both the characterisation and representation. '*Characterisation'* refers to conceptual description, for example the set of features taken to be constitutive of the concept to be measured. '*Representation'* refers to the formal apparatus that gives mathematical or diagrammatic structure and expression (or both) to the measure. '*Procedures'* refers to the 'objectively- observable' procedures for ascertaining values of the measured concept for individual items. 'Objectively-observable' refers to features that are taken to be observable in the actual world such that some set of physical measurement practices – for example those that record preference orderings on surveys or, in the natural sciences, those that measure temperatures on thermometers - can be used to capture those features[27].

Four basic measurement scales are: nominal, ordinal, interval, and ratio. Take an example of a 100-meter race in an athletics tournament. Imagine three runners are participating from three different cities in the UK. Each runner is assigned a number to differentiate from each other. The number displayed to identify runners is an example of *nominal* scale; it provides a label that differentiates them. Once the race is over, the winner is declared along with the declaration of first runner up and second runner up, based on who reaches the destination first, second and third. The rank order of runners such as "third place as 3", "second place as 2" and "first place as 1" is an example of an *ordinal* scale. During the tournament, a judge may be asked to rate each runner on the scale of 1–10 based on certain criteria, such as their running style or strategy. The rating given by the judge employs an *interval* scale. An interval scale measures variables that exist along a common scale at equal intervals. For example, IQ is measured on an interval scale. IQ is numeric data expressed in intervals using a fixed measurement scale. The time spent by each runner in completing the race can be

---

[27] Thank you, Travis Chamberlain, for sharing his work with me on Representational Theory of Measurement. His work informs this section.

represented using a *ratio* scale. Ratio scales have a non-arbitrary zero point. All measures using a ratio scale will have this feature (e.g., 0 denotes absence) even though, for example, metric and imperial have different units.

Empirical facts about the concept that we are trying to capture fix this aspect in a way that social science concepts don't. For example, consider the case of measuring units of length, say with a ruler. The concept of 'length' can be characterised by describing facts about it that hold in the empirical world. Repetitions of that procedure can be counted as units of length represented along a numerical scale in which - because the scale has a zero unit - different units of length are proportional[28].

There is much work to be done before we can arrive at a measurement of something (Bradburn, Cartwright and Fuller, 2016: 3). The measurement of concepts that are useful for policy improvement is a process of continuous refinement as more is learned from each stage (Cartwright and Runhardt, 2014: 265-280). There is a lot of substantive knowledge that informs this process (Cartwright and Runhardt, 2014). But, as I shall argue, substantive knowledge is not always sufficient to settle the design of a measure. This is especially true for social science concepts like my own example of the political commitment to address hunger and undernutrition and even concepts that might seem more objectively fixed like inflation.

4.3 Value judgment in measurement: two examples

I do not defend any particular characterisation of any of the concepts I discuss here. The use of the concepts to be discussed below are for exemplification only. There are, of course, different debates within the three distinct research disciplines mentioned in this chapter. Each discipline has its own different standards. The 'representational theory of measurement' (RTM) is widely shared across the social sciences (e.g., sociology, economics, social policy) and this is why it is used here. What is important to recognise is that choosing one characterisation of a concept rather than another can be very value-laden.

---

[28] Thank you to Travis Chamberlain for sharing his work with me on Representational Theory of Measurement. And this example.

For a great many familiar social science concepts social scientists make choices about which features of the world are important and then construct or model the objects of social scientific knowledge to represent these features. Social science concepts are often vague in some way without being unclear or meaningless altogether. One way to think about these concepts is following Lorraine Daston and Peter Galison who use the term "working objects" (Daston and Galison, 1992: 85). Working objects provide knowledge that will work to achieve a scientific goal. We represent concepts to measure because we believe that they will provide answers to pressing questions. This requires considering what questions are relevant to ask and whether they have cultural significance (Putnam 1981: 202). Another way to think about familiar social science concepts like civil war or inflation is as family resemblance concepts[29] - or what Cartwright following Otto Neurath calls "Ballung" concepts. A Ballung concept applies when there is no one shared feature or essential conditions, but there is significant overlap among the visible features to warrant it being given a category label (Cartwright and Runhardt, 2014: 268).

The example of a *game* is often used as games are connected by overlapping similarities, but it is hard to identify one essential feature all games have in common. There is sufficient overlap among the visible features to warrant the use of the concept, but there is no unequivocal way to display the phenomenon we are interested in (Cartwright and Runhardt, 2014). Ballung concepts are imprecise by nature. However, because we need a more precise concept for social science, we find ways to make the concept more precise to fulfil the job we want it to do (Cartwright and Runhardt, 2014). This usually involves cutting off much of the original meaning thereby losing the ability to serve other purposes.

4.3.1 Example 1: Civil war

Consider the measurement of civil war, as Cartwright and Runhardt do (2014: 266-276). Choosing one characterisation of a concept like civil war rather than another can be very value-laden. When characterising a concept, social scientists must choose whether certain

---

[29] Ludwig Wittgenstien has the best-known exposition of family resemblance concepts in his book *Philosophical Investigations* (1953).

aspects of the category come under the concept, fixing which features belong and which do not (Bradburn, Cartwright and Fuller, 2016: 3), and this depends on the purposes to be served by the concept (Cartwright and Runhardt, 2014: 267).

When seeking to measure civil war, social scientists must first make choices about what to include and where to draw the boundary (Cartwright and Runhardt, 2014). We cannot include everything in a concept of civil war and so we must set boundaries around what is important to measure. Social scientists can be conflicted by competing purposes, such as whether to measure instances of violent conflict in a way that is very accurate to what is going on in a set location or to make cross-country comparisons (Cartwright and Runhardt, 2014: 266). Some scientists might perceive one concept of civil war to be too general for their purposes and put forward a narrower category. We might see a measurement perform very well at a certain job (e.g., context specificity) while performing very badly at others (e.g., comparison across cases). Therefore, we can think about measurement as trading-off between different use purposes (Cartwright & Runhardt, 2014). The particular job that a measure will be used to achieve becomes something against which to judge the appropriateness of the measure for that use.

For example, the concept of *ethnic civil violence* may be used to capture the differences among communities and the power relationships that exist between communities and the state in Cartwright and Runhardt's example of civil war (Cartwright and Runhardt, 2014: 270). Such a measure of civil war is sensitive to ethnicity, but ultimately serves a different purpose to a measure that (for example) focuses on nationalities. The use of a narrower category of ethnic civil violence increases our awareness of ethnic diversity but does so at the expense of monitoring civil war across cases (Cartwright and Runhardt, 2014). We cannot reflect everything, and so a decision to focus on one category rather than another, like the narrower category of ethnic civil violence, prioritises one particular purpose over others (e.g., understanding specifically ethnic civil violence).

When we try to devise measures for concepts that are of interest to us, we may end up distorting what we mean in concrete cases or produce a measure that fails to be fit for the purpose of accurate description (Cartwright and Runhardt, 2014: 276). One tension illustrative of efforts to represent civil war is whether to provide a measure general enough for comparison or to focus on context, as mentioned above (Cartwright and Runhardt, 2014). Civil war can be represented in terms of presence or absence using a two-valued variable, such as 0 or a 1 (Cartwright and Runhardt, 2014: 272). An interval or ratio scale would allow social scientists to label the degree of civil war, for instance in accordance with the number of deaths (Cartwright and Runhardt, 2014: 272). While, on the other hand, a probability distribution would allow social scientists to explore and represent how strongly certain variables correlate with the instance of civil war onset in a state (Cartwright and Runhardt, 2014: 272). If certain features are not included in a standardised measure, important information that is relevant to our understanding of a particular concept or of a particular concept in a particular context can be missed (Cartwright and Runhardt, 2014).

Similarly, there might be cases where an entirely different type of representation would be more suitable. For instance, a probability distribution might offer a particularly useful representation in cases where social scientists are interested in developing causal hypotheses subject to further tests (Cartwright and Runhardt, 2014: 273). Deciding whether to represent civil war using a probability distribution or a ratio or interval or ordinal scale requires the social scientist to think about what it is they want to achieve (Cartwright and Runhardt, 2014). For instance, if social scientists need to offer a ranking, then they will need to use an index and not a vector of unweighted indicators. If social scientists need to compare deaths and violence, for example, then it may be best to use a probability distribution. A social scientist must choose a particular representation of a concept. But when a social scientist adopts a probability distribution, for example, then this can have the knock-on effect of making differences between indicators less obvious.

Sometimes efforts to study something can be restricted by what it is possible to measure (Cartwright and Runhardt, 2014). Social scientists may not be able to gather data on women or other vulnerable groups and so be forced to focus on the data available (Cartwright and Runhardt, 2014: 270). It could be that it is not possible to measure civil war in one way and so there are restrictions on how we can characterise and represent the target concept. In such instances, social scientists may need to return to earlier stages of measurement development to properly explain what is to be meant by civil war in the measure they are devising (Cartwright and Runhardt, 2014). Social scientists may revise the initial understandings of civil war on the basis of results derived from its representation and measurement on the ground (Cartwright, Bradburn and Fuller, 2016: 3). These are choices about the best level of representation that have important (and unavoidable) implications for how we understand civil war in policy settings.

Briefly consider the example of *inequality.* Some indices treat inequality as vertical (e.g., concentration of wealth) while others as horizontal (e.g., exclusion). Vertical inequality is the measure of inequality among individuals or households. While horizontal inequality is seen when people of similar origin or intelligence still do not have equal success and have different status, income, and wealth (Deaton and Paxson, 1998). One key difference is that vertical inequality consists in inequality among individuals or households, but horizontal inequality is defined as inequality among groups and is typically culturally defined (e.g., by ethnicity, religion, or race). Only using vertical inequality as a measure of extreme inequality would render the differences between ethnic groups invisible and so focussing on inequality in these different terms (e.g., vertical or horizontal) warrants different responses. As soon as you settle on a representation of a concept it becomes very easy to lose interest in another way. Neither way is better in any objective sense but those interested in inequality as an issue of wealth concentration among individuals will likely see the vertical indices as more relevant (see also Deaton, 1998 in Reiss, 2017).

If you’re someone that believes that it is important to represent inequality as a concentration of wealth among individuals, then you will use the vertical concept. If you do this, then you will make the differences between groups less visible in your measure. This is because vertical inequality doesn’t focus on the differences between groups.

It could be argued that social scientists could avoid making a choice by providing measures that serve different purposes. This would involve providing a measure of civil war as a ratio scale and providing a measure of civil war as a probability distribution, for example. However, if we offer too many measures for the same concept then this has the effect of making it difficult for policy and public actors to hold measures of civil war to scrutiny (Cartwright and Runhardt, 2014: 275). This could lead to a situation where scientists, politicians or different interest groups cherry-pick measures to suit their own ends (Cartwright and Runhardt, 2014: 275). Or it might push social scientists to try to devise common metrics for central social science concepts. All these issues have the adverse effect of distorting what we mean in concrete cases and fail to be fit for the purpose of accurate description (Cartwright and Runhardt, 2014: 276).

So, we devise measures to suit the purposes that they are intended for. But this can have a knock-on effect of some issues being better represented than others (Cartwright and Runhardt, 2014). Where social scientists fail to capture certain features that are important then this has the knock-on effect of some issues being promoted over others (Cartwright and Runhardt, 2014). Weighing the importance of the various interests that measures could serve is an important task for those social scientists and others who devise measures, who refine their attempts to measure concepts like civil war or violence more broadly as they learn more about the different use purposes (Cartwright and Runhardt, 2014).

Not only are the choices I discuss here *purpose relative* but choosing a purpose can seldom be value-neutral (Cartwright and Runhardt, 2014). Given the uses to which the measure will certainly be put, there will be some harms and some benefits[30].

4.3.2 Example 2: Inflation

Consider the example of *Inflation*[31]. Reiss (2008) considers the construction of *the Consumer Price Index*, which is the most widely used measure of inflation. He demonstrates exactly how values are influencing a concept of inflation and what the value implications are.

---

[30] Choosing between different options is often – and often should be – a matter of deciding what we most value in designing the measure (Reiss, 2013).

[31] Inflation is the quantitative measure of the rate at which the average price of a selected basket of goods increases over time. Inflation statistics form the basis of government policy, which affects firms' and individuals' planning and is used to set wages, welfare benefits and interest rates.

The Consumer Price Index serves as a measure of change in the price of a basket of goods over time and weights the average price of a basket of goods and services, including housing, medical care, food, and beverages, among other things. The Bureau of Labor Statistics (United States) measures Consumer Price Index by a Laspeyres index:

> If q_{i0} denotes the quantity good *i* in the base-period basket at time 0 and p_{i0}, p_{i1} denote its base- and current-period price at times 0 and 1, respectively, then $L_{01} = \sum q_{i0}p_{i1}/\sum q_{i0}p_{i0}$ is the price index for the current period relative to the base period or *Laspeyres price index* (Reiss, 2013: 150).

The *q*s represent the quantity of the good consumed. Which period should we use to fix these quantities if we want to measure how much prices change? There isn't one 'correct' way to do this[32] (Reiss, 2013). The Bureau of Labor Statistics must choose, for example, whether to use the weights for the prices of the goods in the basket of a base period (e.g., a reference point in time used for comparison with other periods) or use the current period quantities consumed (Reiss, 2013: 153). This choice matters if process and quantities change across periods such as when people decide to consume more of some goods and less of others (e.g., more food and beverages and less recreation) (Reiss, 2013: 153). The *Laspeyres price index* weights the basket against a base period, as we see in the quotation above. The Laspeyres is outdated very fast because it "freezes people's tastes in the base period, which will become increasingly unrepresentative" (Reiss, 2013: 153). But it is easier to ascertain values for it because it only requires price observations in the new period. There is an alternative index, called *the Paasche index, that* uses weights from the current period and so is more up to date (Reiss, 2013: 153). But, following Reiss, this is harder to measure because it requires quantity observations in the new period in addition to price observations because it uses current period weights.

Whether to use a Laspeyres index or a Paasche index or some other index shouldn't be thought of as a factual judgment because choosing one over the other means that we commit to certain things. For instance, we must decide whether the adjustment that the Paasche index allows is worth the additional cost of measuring it.

---

[32] See also Dupre (2007) for a similar construction of this argument.

A traditional ‘fixed-basket’ index tracks the cost of purchasing a given basket of goods over time. The Laspeyres price index is an example. It answers how much a consumer’s income must change between the base and a current period to be able to buy the designated basket of goods (Reiss, 2008: 44). But recall that in a Laspeyres formula the quantity weights are fixed in the base period rather than the current one. With the adjusted price one can still buy the same basket this year as last, but it may be possible to buy more, or less, utility with that amount of money. That’s because when relative processes change (e.g., new items appear in shops), some items in the basket become harder to find and thus more expensive or items in the basket improve or diminish in quality without a price change for them. So, for example, consumers can optimise by switching from one more expensive item to a cheaper item of the same utility and thus obtain more utility from the amount of money it costs to buy the designated basket (Reiss, 2008: 44).

Because of this a Cost-of-Living Index is widely accepted among economists as opposed to ‘fixed-basket’ indices like the Laspeyres. The Cost-of-Living Index “measures the cost of purchasing a given amount of ‘utility’ or ‘welfare’ or ‘standard of living’ (in discussions about the Consumer Price Index, these terms are used interchangeably)” (Reiss, 2008: 39). But developing a Cost-of-Living Index requires making a lot of choices. Not all these choices are easily understandable to outsiders, and many are not within the normal expertise of those designing the measure (Reiss, 2008: 39). One problem is that statistical agencies do not know the utility attached to each of the goods in the basket, “hence, a ‘true’ Cost-of-Living Index must be approximated on the basis of assumptions about the underlying utility functions of the customers” (2008: 27). The assumption behind the Boskin Commission’s usual attributions of utility was that the consumer only substitutes goods on the basis of their cost and therefore implies that item substitution is caused by no other factor than price changes. But consumers also substitute goods because of a range of other factors, such as availability, chance, access, finances, or many more (Reiss, 2008: 27).

Another choice is how to deal with "quality changes" (Reiss, 2013: 155). For instance, Reiss asks, "how should we deal with the fact that a new good differs in quality from an old one?" (Reiss, 2013: 155). He continues,

> If the new good constitutes an improvement over the old one and it is treated as equivalent substitute, inflation would be over- estimated because consumers profit from better goods and a higher price is justified because companies invest in their development. On the other hand, if price changes are interpreted as reflecting nothing but quality changes…inflation would be underestimated because price changes are often introduced at the same time as quality changes in order to hide them (Reiss, 2013: 155).

Price cannot be considered as having inflated if we are getting a better product at a higher price and so it is not possible to measure inflation solely by an increase in price (Reiss 2013: 155). Whether or not something is a better-quality product is a question about what is valuable to a particular person or group. Reiss notes that we could approach the question of whether the new good constitutes an improvement over the old one in factual terms using an economic approach to index numbers (Reiss, 2013: 155). That approach supposes that how much a person values a good is reflected in their choices about buying that good. But this approach often does not provide correct information about how much a consumer prefers one product over another since, for instance, consumers might want to continue making the same purchases that they did in the base period but are not able to due to price changes (2013: 154). The choice of approach to use is a subtle value choice that needs to be made in the design of a measure that will have different effects (Reiss, 2013).

The procedure for a Cost-of-Living Index is approximated on the basis of assumptions about the underlying utility functions of the customers (2008: 27). Recall that statistical agencies do not know the utility attached to each of the goods in the basket. So, one assumption is about the substitution behaviour of consumers (Reiss, 2008: 27). Making these assumptions requires real commitment on the part of the investigator to regard certain facts as relevant to the measurement procedure (Reiss, 2008). So, statements about a cost-of-living inevitably reflect value judgments (e.g., whether consumers actually benefit from new goods) (Reiss, 2013).

This is a significant claim. There is a value judgment in whether a good is *better* or not. It is impossible to know whether a good is actually better because the value of the product is different for everyone. Consider the simple example of bread. Let's say bread used to be $1 and now it's $2, yet it does have fortified grains and seeds (e.g., it is a *better* product). However, it is now double the price. Whether we can really say that this is an equivalent product cannot be done without making the assumption that the improvement in quality is irrelevant and the value placed on the bread by the consumer is the same (e.g., the $1 bread that existed no longer exists and there is a better $2 bread in its place). Whether the new bread is better than the old version is still a value judgment: would the consumer want the new bread? The unavoidability of these questions shows a clear space where value judgment plays a key role in setting the Consumer Price Index that has real tangible potentially negative impact on the policies that the Consumer Price Index can be used to influence.

Following Reiss and others, the assumption that we can offer purely factual descriptions of economic concepts (e.g., a Cost-of-Living) and leave others to decide what to do with it may not be possible (see Dupre, 2007: 36). We can offer conceptual statements about concepts significant to policy, like inflation. But if we want to assess whether the procedures used to offer the statement are the "right" ones then this cannot be assessed without value judgment.

Given different measurement purposes, an account of inflation will perform better or worse. If we are to use the measure of inflation to adjust people's eligibility for different types of government assistance (e.g., social security, military, and federal civil services retirees) then it seems desirable to use a measure that prices goods in the shops accessible to them and not in retail outlets they can't get to. But this may result in a different Consumer Price Index. If you take into account that those on benefits often don't have access to personal mobility, then they are much more likely to shop in smaller local shops where the prices are higher.

The use of Consumer Price Index to adjust people's eligibility for different types of government assistance means that if the Consumer Price Index takes account of the price of local goods, then this advantages the benefit recipients and disadvantages the taxpayers who pay for those benefits. Whether the concept of inflation is “adequate” can’t be known independently from the purpose for which it is to be constructed (Dupre, 2007: 38). Reiss argues that many of the choices in the construction of an index (e.g., Consumer Price Index) are value-laden because they advantage, or disadvantage, certain groups given a particular purpose to which the Index is to be put. Given the purpose of using the Consumer Price Index to set pension benefits that the taxpayer pays for, doing it one way over the other advantages the interests of one of these groups vs the other (Reiss, 2008).

One of the issues in more recent years, as highlighted by Cartwright and Runhardt (following Reiss, 2008), is that people have started shopping in discount shopping centres rather than more traditional outlets. The Boskin Commission (United States)[33] observed this gain in discount stores and decided to adjust the Consumer Price Index, which affects many other things that are pegged to the Consumer Price Index, like pension payments and various kinds of benefits. But the Consumer Price Index fails to capture certain features that are important which then has the knock-on effect of some issues being better represented than others (Cartwright and Runhardt, 2014). In this case, the shopping patterns of the elderly who do not use the discount stores are not reflected, which can lead to negative effects on them if their pensions are affected by the adjustment to the Consumer Price Index. This is a particular type of information that is missing and has a knock-on effect of marginalising vulnerable groups (Cartwright and Runhardt, 2014). Cartwright and Runhardt argue that there is a lot of substantive knowledge (and value judgments too, as I discuss in the next section), such as in this case knowledge of shopping habits of different groups, that should inform the process of designing a measure (Cartwright and Runhardt, 2014).

Reiss’ point is not only that decisions about retail stores show the design of the Consumer Price Index to be purpose relative but that choosing a purpose can seldom be value-neutral.

---

[33] An independent panel of experts (the Boskin Commission) were appointed by the US Senate Finance Committee to study the accuracy of the standard measure of inflation, the Consumer Price Index.

Given the uses to which the measure will certainly be put, there will be some harms and some benefits. Where the choice of purpose is not already dictated it is surely morally irresponsible not to make reasoned value judgments weighting harms and benefits. And sometimes where the purpose is specified (e.g., social scientists are commissioned to deliver a measure of F for purpose P) there are cases where the scientists ought to hesitate at doing so if it is predictable that, given the predictable uses for the measure, one that is good at measuring F for purpose P will be good at that but when used will cause some side substantial harm. For example, as Reiss outlines, economists pay special attention to benefit recipients, which means that the taxpayers suffer. So maybe a social scientist should urge that this kind of measure is not devised. It is in this way that choosing among purposes is usually value-laden.

4.4 A new example: The Hunger and Nutrition Commitment Index (HANCI)

Finally, I offer a new example of the Hunger and Nutrition Commitment Index (HANCI) produced by researchers at the Institute of Development Studies (United Kingdom). This is an international comparison tool, which ranks governments on the level of political commitment to address hunger and undernutrition. The index is intended for use by nutrition stakeholders as a tool for political advocacy (te Lintelo et al. 2013). The first HANCI report in 2013 clearly set out its intended use:

> (a) by credibly measuring commitment it will strengthen our ability to hold governments to account for their efforts in reducing undernutrition and hunger; (b) if civil society is better able to hold governments to account, it can apply pressure and ensure that hunger and undernutrition are put high on development agendas; (c) governments can hold themselves to account in their efforts to keep hunger and undernutrition high on the agenda: the index can help them to track and prioritise their efforts because the index is constructed on the basis of performance in different areas (legal, policy and expenditure); and (d) commitment can be linked to outcomes, to allow all to assess the 'value added' of different commitments and effort. (te Lintelo et al. 2013: 2).

I will first offer an overview of work based on empirical work by researchers at the Institute of Development Studies (te Lintelo et al. 2013; te Lintelo & Lakshman, 2015; te Lintelo, 2019). This is primarily the work of lead researcher Dolf te Lintelo. I use this material and indicate where I have done so. My goal in reviewing this material is to demonstrate that measurement choices are value-laden. I will go beyond the original material presented by the researchers and hypothesise ways in which measurement choices can have a possible negative effect on some policy interests over others. I can’t claim how likely these effects are as this would require a lot of empirical data. My point is only to demonstrate that there can be consequences. There can then be the issue of who should make the value judgment, but this is beyond the scope of this dissertation.

I have an insight into this example as I was part of the evaluation team for HANCI and was a contributor to the most recent report on HANCI, ‘Process Tracing the Policy Impact of ‘Indicators’’ (te Lintelo et al. 2020). However, this material does not form the basis of my discussion. The purpose of the evaluation was to use a process tracing methodology to understand the impact of HANCI indicators on nutrition policy in various country contexts. Whereas my focus here is to understand the role that values play in the very construction of the index.

Political commitment has been a feature of mainstream international development policy since the 1990s (Hammergren, 1998: 12). Yet it is considered as one of the slipperiest concepts since it is “never defined except by its absence” (Hammergren, 1998: 12). Another feature of political commitment is that it does not exist without aid. Rather, it is something that is created or strengthened by certain types of strategic action (Barker et al. 2018). According to Ypi (2016: 603), “to be politically committed means to care about the public good and actively seek to promote it, making one’s efforts and ideas of social change part of a joint project shared with others”. There is a vast literature to support political commitment as a worthy measure to understand the problem of hunger and undernutrition (IFPRI, 2014; te Lintelo et al. 2013; Gillespie et al. 2013; Hammergren, 1998). Metrics and scorecards of political commitment have proliferated in different disciplines, including food security (te Lintelo & Lakshman, 2015). Yet despite the mounting interest in the measurement of political commitment to address hunger and undernutrition, there is no agreement on the best way to achieve it.

HANCI is an interesting example because it is both a measurement of political commitment *and* a tool for political advocacy. There are relatively few tools that are built for the purpose of measuring political commitment and political advocacy (Baker et al. 2018; Fox et al. 2015). HANCI is distinctive in its ambition to measure political commitment while developing and sustaining a momentum for it (et Lintelo et al. 2020; Gillespie et al. 2013).

What is particularly interesting is that HANCI is informed by a co-production approach and the researchers form a partnership between social scientists and nutrition stakeholders working on the ground[34]. Working in partnership with nutrition stakeholders has the benefit of utilising knowledge from actors both within and outside academia. What this means in practice is that stakeholders were involved alongside social scientists in significant choices: such as, the selection of indicators and the weighting of those indicators. Practically, this may improve the likelihood of the index being used to inform policy (Gillespie et al. 2013). Normatively, we can ask what biases – potentially in favour of larger, well connected organisations – this introduces into the policy process.

This is an example where the researchers are trying to limit the influence of their value judgments by bringing stakeholders into the design process. The researchers have attempted to remove as much as possible their value judgments and approached the design of the index with self-awareness while attempting to shift value judgment to stakeholders. There is room to consider the role that values are playing. My chief concern is to make values explicit so there is at least opportunity to critically examine them.

4.5 Four value-laden design choices

I outline four subtle value-laden choices in the design of the HANCI that concern how to display indicators and how to determine the importance of various indicators. The first is the choice to display indicators in a uniform range. The second choice is to use a ranking

---

[34] For example, communities affected by undernutrition, civil society organisations representing the interests of those affected by undernutrition, leading interest groups, including the Scaling Up Nutrition civil society movement.

scheme. The third choice is the scale for the ranking. The fourth is the choice of weighting scheme.

Table 1: HANCI indicators weighted by theme

Laws **(e.g., legal frameworks – for example the level of constitutional protection of the right to food)**	**Policies** **(e.g., government programmes and policies – for example the extent to which nutrition features in a national development policies/strategies)**	**Spending** **(e.g., public expenditures – for example the percentage of government budgets spent on agriculture)**
Level of constitutional protection of the right to food	Access to land (security of tenure)	Public spending on agriculture as a share of total public spending
Equality of women's access to agricultural land	Access to agricultural research and extension services	Spending on health as a share of total public spending
Equality of women's economic rights	Civil registration system (coverage of live births)	Separate budget for nutrition
Constitutional right to social security	Functioning of social protection systems	
Enshrine ICBMS In domestic law (The extent to which the International Code for Marketing of Breastmilk Substitutes is enshrined in law.)	Vitamin A supplementation coverage for children	
	Government promotes complementary feeding	
	Population with access to an improved water source	
	Population with access to improved sanitation	
	Healthcare visits for pregnant women	
	Nutrition features in national development policy	

	National nutrition policy / strategy	
	Multi Sector and multi stakeholder policy coordination	
	Time bound nutrition targets	
	National nutrition survey in the last 3 years	

HANCI, 2020

Before turning to the first design choice, consider the selection of the 22 indicators[35]. HANCI compares and ranks the performance of 45 ‘high-burden’ countries based on 22 indicators of political commitment (te Lintelo et al. 2013). The indicators are grouped under three themes,

> Laws (Legal frameworks – for example the level of constitutional protection of the right to food)
> Policies (government programmes and policies – for example the extent to which nutrition features in national development policies/strategies)
> Spending (public expenditures – for example the percentage of government budgets spent on agriculture) (HANCI, 2021)

The indicators represent different types of hunger and nutrition interventions, ranging from vitamin A coverage, complementary feeding, skilled health attendance for pregnancy women, access to agricultural extension services, multi-sectoral and multi-stakeholder coordination mechanisms, as well as indicators relevant to wider political and financial enabling environments (te Lintelo et al. 2013: 7). The indicators include a wide range of measures, including what have been termed ‘expressed commitments’ as well as ‘actual budgetary commitments’ (te Lintelo et al. 2013). An expressed commitment, for example, reflects plans communicated in parliamentary debates or other speech acts. An actual budgetary commitment, on the other hand, is the amount of spending a government is planning to take in a given period (e.g., the fiscal year) in a particular area of undernutrition.

---

[35] Table 1 displays 10 indicators relevant to addressing hunger reduction and 12 indicators relevant to addressing undernutrition.

This is often considered to be a particularly useful area for mobilising advocacy groups, as it serves as a point of reference to hold governments accountable (te Lintelo et al. 2013).

4.5.1 The choice to display indicators in a uniform range

The first design choice I shall look at involves normalising and aggregating the indicators (te Lintelo et al. 2013: 16). The specific aggregation issue having to do with weights will be dealt with separately. The indicators were aggregated at the theme level, where the themes include, legal frameworks (5 indicators); policies and programmes (14 indicators) and public expenditures (3 indicators), as displayed in Table 1. The indicators needed to be normalised before being aggregated to prevent any of the indicators having too much weight in the index because of the unit of analysis (te Lintelo et al. 2013: 16). The normalisation procedure used was adopted from the United Nations Development Programme (UNDP) Human Development Index (HDI), which rescales indicators using the formula outlined in te Lintelo et al (2013: 16),

$$\text{normalised value of } xi = x\ddot{o}i = \frac{xi - xmin}{x_{max} - x_{min}}$$

This normalises values within a [0,1] range. This type of normalisation was important to fulfil the purpose of displaying indicators in a uniform range. Normalising the indicators within a [0,1] range is good for comparing indicators more easily. This is useful because there are many different types of indicators, and the normalisation avoids skewing the data in favour of any one indicator.

Take a simple hypothetical example. Imagine Tanzania has a total public spending of 100 million and a health spending of 25 million. Within a [0,1] range this would be represented as 0.25. Imagine Nigeria has a total public spending of 75 million and health spending of 24 million. Within a [0,1] range this would be represented as 0.32. So, even though Nigeria actually spends less on health than Tanzania, the score is higher. What the researchers have done with the normalisation is remove the possibility of a country appearing to spend significantly more than other countries on hunger and undernutrition because countries are assessed against their own context before being compared to other countries. Hence, normalising the indicators within this range avoids skewing the data.

An alternative normalisation method might include a ‘standardisation’ (or Z-score) (te Lintelo at al. 2013: 16). A Z-score would be useful for showing deviation from an average. It is frequently employed in undernutrition measures to show the extent to which a population differs in weight from an average point. So, it can do a good job of representing cases that fall below (or above) an average. But a Z-score was not selected because the researchers wanted to display indicators in a uniform range. The choice of standardisation technique is an advantage or a disadvantage depending on what we want to achieve with the measurement.

4.5.2 The choice to use a ranking scheme

A second choice is the use of a ranking scheme, which has implications for how political commitment is displayed across the 45 countries included in the index. What makes the use of a ranked metric like HANCI particularly attractive is that the constituent units can be compared and ranked according to some criteria across contexts. Such numeric and simple information is intended to serve as an advocacy tool for use by nutrition actors to convince governments and the general public of the need for action (te Lintelo et al. 2013: 2). There are advantages and disadvantages to offering a ranking to display the data. Whether or not a ranking is suitable for the construction of the index has to be considered against the intended use. For example, if the index is intended to demonstrate performance as a simple comparison, then given this particular intended use, a ranking of 45 countries is appropriate. But alternatively, if an index is intended to serve as a tool to offer context-specific information on different countries and their achievements, then a ranking is less appropriate.

HANCI seems to presuppose that a comparison between countries is a valuable strategy for convincing the government and general public of the need for action. This is not a value-neutral assumption and there can be disadvantages to displaying data in this format.

One disadvantage is the loss of context-specific information. For example, context-specific information on the level of commitment at the district-level is not represented in a ranking of 45 countries (te Lintelo, 2019). A district-level comparison would need the operationalising of a different set of indicators (te Lintelo, 2019: 8). This may include budget spend on nutrition as a share of total district spending (te Lintelo, 2019: 8). Or inter-district differences in nutrition outcomes (te Lintelo, 2019: 8). In countries like Tanzania 'spending on health as a share of total public spending' (as per the indicator in Table 1) does not translate to spending on health at the district-level. This is because money is disbursed by central government authorities, though it often does not reach local government and/or undernutrition services at the community level.

Displaying data as a ranking of 45 countries may have the consequence of masking this particular issue in policy debates. There is a risk that the level of political commitment within a particular country may be overestimated. This has the consequence of making it look like a government is performing better than it actually is, making it more difficult for specific districts to hold national governments accountable to their commitments. For example,

> In Tanzania, over the past fifteen years, *nutritional status improved across Mainland Tanzania* (excluding data for Zanzibar). Tanzania Demographic Health Survey (TDHS) data shows that stunting of children under five years of age in mainland Tanzania dropped from 48.4 per cent in 1999 to 34.4 per cent in 2015. *However, stunting prevalence varies significantly by zone, and by region (districts are a subdivision of regions).* The Central and Eastern zones reduced stunting by nearly 50 per cent between 1999 and 2015, compared to only 12 per cent and 13 percent declines in Lake and Northern zones, respectively (te Lintelo, 2019: 7, emphasis added).

The use of a ranking to display data emphasises comparison between countries. But it may do so at the expense of a better way to examine context-specificities within a particular country. It is likely that countries will have differing needs and priorities to consider depending upon their specific population and the root causes of hunger and undernutrition within a particular context. For example, in countries with high dependence on food imports and low domestic production, nutrition education, consumer protection and financial accessibility to nutritious food that is not overly processed may be key. While, in predominantly rural areas, where only limited types of food may be available locally, issues such as access to water and sanitation, school food and nutrition and economic access to adequate quantity and quality or diversity of foods may be important areas for legislation. Whereas, in urban areas, food safety and quality, labelling, food composition - for example, salts, sugars, and fats in processed food - and marketing may be priorities. In such cases, a lack of context-specific information is a disadvantage.

The choice to use a ranking facilitates an international comparison. But this puts out of focus other political issues within a particular country that are relevant to political commitment. It may be assumed that an international comparison is useful for promoting accountability, but this can have the knock-on effect of making other issues less visible in policy debates.

The choice to use a ranking is not value-neutral. If researchers overestimate the level of commitment to address hunger and undernutrition a country actually has, then this may negatively impact communities suffering from undernutrition on the ground. As soon as researchers settle on a representation of political commitment it becomes very easy to lose interest in another way. Often neither way is better in any objective sense.

There is an option of providing measures that serve different purposes. This would involve providing a measure for district-level comparisons and providing a measure of international comparisons, for example. But, if we offer too many measures for the same concept then this has the effect of making it difficult for policy and public actors to hold measures of political commitment to scrutiny (Cartwright and Runhardt, 2014: 275). This could lead to a situation where nutrition stakeholders cherry-pick measures to suit their own ends (Cartwright and Runhardt, 2014: 275).

4.5.3 The choice of measurement scale adopted for the ranking

A third choice is the measurement scale adopted for the ranking. There are a variety of measurement scales and supporting procedures to choose from. But there are two standard choices: an ordinal or a cardinal scale.

A cardinal scale is useful for representing how much better one thing is than another. This could be constructed by taking a total of absolute values across the three themes: 'legal frameworks' (for example the level of constitutional protection of the right to food), 'policies and programmes' (for example the extent to which nutrition features in national development policies/strategies), and 'public expenditures' (for example the percentage of government budgets spent on agriculture) (te Lintelo et al. 2013: 17). One disadvantage of the use of a cardinal scale is that it allows outliers to dominate the aggregated values. An ordinal scale was chosen by the researchers as it tells the position of something in a list, like first, or second, or tenth, without any commitment about the size of the gaps. Ordinal data are discrete integers that can be ranked or sorted. The use of an ordinal scale allows the adding up of the ranks achieved for each theme where the country that achieves the lowest sum of ranks is shown to perform the best (te Lintelo et al. 2013: 17). One advantage of an ordinal scale is that it facilitates comparison of scores across the 45 countries without a commitment as to how *much* better one is than another. Ordinal scales have real advantages for end-users. For example, they give the end-users a 'stick' to hold people to account with. However, a defining characteristic of ordinal data is that the distance between any two numbers is not known.

The distance between first and second may not be the same as the distance between second and third. Consider again the well-known example of a race. The winner might have completed the race in 35 minutes and 0 seconds, second place might have completed in 35 minutes and 1 second, third place might have completed in 50 minutes and 1 second. The order is what matters, but the differences between the orders does not matter. Without additional data, we do not know the relative distance between the ranks. We do not know by how much the first country in the ranking commits to targeting undernutrition over the remaining countries in the ranking. And so, without additional information, the index number assigned by HANCI may be misleading, or lead to incorrect assumptions, about how much political commitment there really is.

One disadvantage of ordinal scales is the loss of information. Statistical measurements require variables, but not all variables are the same. Recall that something like length has a non-arbitrary '0' point. This is what Patrick Suppes calls an 'empirical constraint'. Rankings of political commitment, though, are a different matter. Ordinal rankings are designed to infer conclusions -- for example, how well a country performs in terms of political commitment against another country. The conclusion attempts to see beyond the actual number to some qualitative conclusion. For example, the rankings can be used to hold countries to account for dropping three places. Or can be used to generate statements like, “South Africa is at the top of the ranking, while Zimbabwe has moved up three places”. “South Africa retains the top-spot, thanks to stable performance”. “Malawi remains in second overall position”. “Other fast improvers include Zimbabwe (+14 ranks); Sudan (+13) and Niger and Cameroon (+10)” (Hunger and Nutrition Index Africa, 2020). Such statements offer useful advocacy messages for nutrition stakeholders working within country and international nutrition movements more broadly.

However, the top 10 spaces may be so finely spaced that they do not actually represent a change in absolute value. So, these are only useful statements if nutrition stakeholders using the data know something about how noisy[36] the ranking is. This might include any data that nutrition stakeholders (or other end-users) cannot understand or interpret correctly, such as the distance between the rankings. End-users require additional information about the ordinal ranking if it is to be meaningful. In order to answer relevant questions, such as “are countries rankings changing?” then we need to understand whether the rankings move for small changes in values and how finely spaced the rankings are[37].

Consider a brief hypothetical example. It could be the case that a particular country does not have ‘the constitutional protection of the right to food’ established in their legal system, as displayed in Table 1. Hence, a particular country scores O for this indicator. However, a different country *without* this basic commitment could end up higher in a ranking than a country with this basic legal requirement. This is due to the fact that they have paid sufficient attention to the other indicators as put forward by HANCI.

---

[36] Data with a large amount of additional meaningless information in it called “noise”.

[37] An ordinal scale may be demotivating for those governments that see their commitment differently. Or who only fall short of the top spot by a very small margin.

4.5.4 The choice of weighting scheme

A fourth choice is how to assign the weights in constructing a single index number from numbers assigned different indicator or indicator categories**.** The weights assigned to the different dimensions fix their relative influence on the final outcome index number (e.g., the ranking) and relative significance of each dimension to the others (Decancq and Lugo, 2010 in te Lintelo at al. 2013).

The following choices were taken in the construction of the index: (1) to use an equally weighted aggregation of indicators by theme and (2) to use an equally weighted aggregation of themes (te Lintelo et al 2013: 16). This means that the three themes (e.g., legal frameworks, policies and programmes, and public expenditure) are treated as having equal weight in reflecting political commitment towards addressing undernutrition (within the sub-indices and consequently in the overall HANCI) (te Lintelo et al. 2013: 13, 18). There are advantages and drawbacks to different weightings and any weighting scheme involves trade-offs between the dimensions of the index (te Lintelo et al. 2013: 17; Decancq and Lugo, 2010 in te Lintelo et al. 2013). A weighting for something like political commitment can be constructed in many different ways and the way in which the weights are assigned will inevitably lend support to one set of issues over another.

The researchers argue,

> While we suggest a trade-off between legal frameworks, policies and programmes and public expenditures, we cannot reasonably uphold this position at the indicator level. *The unequal weighting of indicators means that for instance, putting the ICBMS into law, or having a nutrition budget is weighted ten times more than coverage of access to sanitation; clearly this is contestable.* Nevertheless, we decided to privilege comprehensiveness over equality of weighting for indicators. That is, we do not want equal indicator weighting to drive down the number of indicators to the lowest common denominator, as we want to capture the multi-dimensional nature of political commitment to reduce hunger and undernutrition (te Lintelo et al. 2013: 18, emphasis added).

What the researchers are acknowledging is that the indicator reflecting 'The extent to which the International Code for Marketing of Breastmilk Substitutes is enshrined in law' is worth *significantly* more than the indicator reflecting 'Population with access to improved sanitation'. Of course, some value perspectives would view access to sanitation as more important than buy-in to the International Code for Marketing of Breastmilk Substitutes. So, the decision to weight by theme has the potential to negatively impact on the representation of certain issues since different value perspectives prioritise things differently.

A similar concern involves the substitutability of the various dimensions of the index (Decancq and Lugo, 2010 in te Lintelo et al. 2013). HANCI assumes full substitutability between the themes of the index (e.g., policies and programmes, laws, and spending). Some assumptions have to be made on the degree of substitutability. The researchers continue,

> Although equal weighting schemes are often defended from an agnostic perspective, they are not uncontroversial. Like any other weighting scheme, equal weighting involves value choices regarding the *substitutability* of various dimensions of the index, without specifying the normative attractiveness of such choices (Decancq and Lugo 2010 in te Lintelo et al. 2013: 18, emphasis added).

There is no reason to believe that the degree of substitutability between 'policies and programmes' and 'public expenditures' is exactly the same as the degree of substitutability between 'policies and programmes' and 'legal frameworks' (Decancq and Lugo, 2010). Whether we can really say that the themes are equivalent requires a commitment to the assumption that no theme counts more than another. There is room for reasoned judgment about the potential harms to certain values and interests of the choice to use an equally weighted aggregation of indicators by theme that assumes full substitutability between themes.

The researchers recognised that the choice of weighting has implications for whose priorities are captured in those weightings (te Lintelo et al. 2013: 16). They devised a simple exercise that allowed an opportunity for others to set their own subjective weights. This is an example where the researchers are trying to limit the influence of their value judgments by bringing stakeholders into the design process.

> In order to compare our subjective allocation of equal weights to themes with alternative preferences, we devised a simple exercise that allows others such as in-country experts, hunger- and undernutrition-affected communities, and third parties to set their own subjective weights. This exercise thus avoids imposing the weighting preference of a group of researchers at IDS (te Lintelo et al. 2013: 18, emphasis added).

If a nutrition stakeholder takes public expenditures to be of central importance with respect to the aims that they have in view for use of the measure, for example, then they can weigh this theme more heavily. This would have the implication of potentially changing a country's position in the global ranking. Or, if a nutrition stakeholder is interested in policies and programmes, for example, they can weigh this theme more heavily according to their own subjective preferences. This is a unique design feature selected by the researchers that allows stakeholders the opportunity to inform the weighting and thus decide what is most important for monitoring political commitment to addressing hunger and undernutrition (te Lintelo et al. 2013: 18).

One concern however is that the selection of stakeholders may lead to the under-representation of key stakeholder groups, which may result in a skewed weighting (Decancq and Lugo, 2010 in te Lintelo at al. 2013). The use of nutrition stakeholders can be advantageous where those who understand hunger and undernutrition can be utilised as specialists. But, if stakeholders from leading interest groups serve in disproportionately large numbers, then others (perhaps minority or other individuals) may be missed.

Normative implications

One possible implication is the biased inclusion of nutrition stakeholders. For example, leading interest groups affiliated with the progress towards food and nutrition security were included in the design of the metrics. The Scaling Up Nutrition Movement was consulted in focus group discussions in various countries (te Lintelo et al. 2013). This movement also includes members of the Transform Nutrition Movement, the Partnership for Nutrition in Tanzania (PANITA), among others from academia (te Lintelo et al. 2013).

Consider a specific example. A key stakeholder is the Scaling Up Nutrition (SUN) civil society network (te Lintelo et al. 2020). SUN advocates for increased investments for the scaling up of proven and context-specific, cost-effective, nutrition-sensitive, and nutrition-specific interventions. They utilise advocacy, responsible nutrition governance and private sector engagement (SUN Movement, 2015). SUN offers countries, including many African countries, direct assistance with policies and projects that bring together governments, civil society, and the private sector for improved nutrition (SUN Movement, 2015). SUN highlights the need for multi-sectoral approaches, coordination and improved accountability and governance. What these groups have in common is a commitment to a collective action strategy to build and sustain political commitment.

An important feature of political commitment is its preference for the *collective* rather than the individual (White and Ypi, 2010 in Ypi, 2016). Collective action differs from more narrow interest groups because the appeal is to the whole rather than only one subsection of the public (Ypi, 2016). Collective action represents a normative aim that articulates how power ought to be exercised and in what way political institutions should enable social cooperation (White and Ypi, 2010 in Ypi, 2016). The belief in the value of collective action disposes individuals to act in concert with others and to sacrifice their short-term interests on behalf of ideals of social change that will serve future generations, even when the outcome of collective action does not directly benefit them personally (Ypi, 2016). Seen in this way, political commitment is enhanced when agents committed to particular political goals form a lasting association that coordinates future action (Ypi, 2016). Strategies that support political commitment inevitably reflect the use of value judgment in that they require commitment to certain beliefs about politics, namely, the preference for the collective rather than individual.

Practically, the inclusion of leading interest groups in the design of the metrics and selection and weighing of the indicators may improve the likelihood of the index being used to inform policy (Gillespie et al. 2013). Normatively, we should ask what biases – potentially in favour of larger, well connected organisations – this introduces into the policy process.

Whereas collective action is often preferred by those seeking to obtain a common goal, individual action is typically preferred by minority group members (e.g., ethnic minority groups, elderly groups) or perhaps those requiring immediate action (e.g., those affected by immediate vulnerability). High-risk groups that are especially vulnerable to undernutrition might include poor farmers, ethnic minorities, geographically marginalised communities, and individuals living in disaster zones. This small-minority may not be included in redirecting resources towards the state, sacrificing short-term interests in favour of future generations. The representation of political commitment as a collective action strategy might mean that nutrition interventions become less available to minority groups as represented by the index, as Reiss outlines.

One implication is that weighing the importance of the various interests is an important task for the researchers and others who devise measures (Cartwright and Runhardt, 2014), who refine their attempts to measure concepts like political commitment as they learn more about the different intended uses. For example, if we characterise political commitment in terms of X (here collective action) and this is more effective overall than Y (narrow interest groups), but it leaves out a significant minority that Y catches, which outcome is valued more—the positive effect on the many or the really negative effect on the few? This moves from the putatively factual issue of which is more effective to an issue that needs values to settle.

Conclusion

There is a lot of substantive knowledge that informs the process of designing a measure (Cartwright and Runhardt, 2014). Substantive knowledge is however insufficient to settle the design of a measure in my examples and instead values play a central role (Cartwright and Runhardt, 2014; Reiss, 2008, 2013).

HANCI is particularly useful as an example as it draws on the expertise of nutrition actors to select and weight the indicators informing the index (te Lintelo et al. 2013). Practically, this may improve the likelihood that nutrition actors use the data. Normatively, we should ask what biases – potentially in favour of larger, well connected, and professionalised organisations – this introduces into the policy process. Given the significance of HANCI for informing policy, social scientists producing the metrics ought to consider places in the construction of an index where we may end up lending support to certain interests over others, and what the possible harms of doing so might be. There can then be the issue of who should make the value judgment, which is beyond the scope of this dissertation. Nonetheless, empirical research is necessary to evaluate the ways in which measurement tools like HANCI influence policy processes: for better and for worse.

References

Bradburn, Cartwright, N. and Fuller (2016) A theory of measurement, Working Paper, Centre for Humanities Engaging Social Science and Society (CHESS), Durham.

Cartwright, N. and Runhardt, R. (2014) Measurement, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 31-47.

Cartwright, N. (2008) *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Daston, L. and Gallison, P. (2007) *Objectivity*, Zone books: London.

Deaton, A. and Paxson (1998) Aging and Inequality in Income and Health, American Economic Review: Papers and Proceeding, 88, (2), 288-253.

Decancq, K. and Lugo, M. A. (2010) Weights in Multidimensional Indices of Wellbeing: An Overview, Leuven: Katholieke Universiteit Leuven/Oxford: Dept. of Economics, University of Oxford, 30 August 2010.

Gillespie, S., Haddad, L., Mannar, V., Menon, P., & Nisbett, N. (2013). The politics of reducing malnutrition: Building commitment and accelerating impact. The Lancet, 382(9891), 552–569. http:// dx.doi.org/10.1016/S0140-6736(13)60842-9.

Hammergren, L. (1998). Political will, constituency building, and public support in rule of law programs (No. PN-ACD-023). Center for Democracy and Governance, USAID.

Hunger and Nutrition Index (2021), *Hunger and Nutrition Commitment Index*, Institute of Development Studies. Available at: http://www.hancindex.org (Accessed: 28th June 2021).

IFPRI. (2014). Global nutrition report 2014: Actions and accountability to accelerate the world’s progress on nutrition. Washington: International Food Policy Research Institute.

Larner, W., and R. Le Heron. 2004. Global Benchmarking: Participating ‘at a Distance’ in the Globaliz- ing Economy. In *Global Governmentality: Governing International Spaces*, ed. W. Larner and W. Walters, 212–232. London: Routledge.

Merry, S.E. 2011. Measuring the World Indicators, Human Rights, and Global Governance. *Current Anthropology* 52: 83–95.

Putnam, H. (1981) *Reason, Truth and History*, Cambridge University Press: Cambridge.

Reiss, J. (2008) Methodology and expert judgment in evidence-based economics: lessons from the CPI controversy, Error in Economics, 1st Edition, London: Routledge.

Reiss, J. (2013) *Philosophy of Economics: a contemporary introduction*, London: Routledge.

Reiss, J. (2016) *Error in Economics: Towards a More Evidence-Based Methodology*, London: Routledge.

SUN Movement (2015) *SUN Civil Society Network*. Available at: https://scalingupnutrition.org/sun-supporters/sun-civil-society-network/ (Accessed 26th June 2021).

te Lintelo, D.J.H. Haddad, L., Lakshman, R. and Gatellier, K. (2013) The Hunger And Nutrition Commitment Index (HANCI 2012) Measuring the Political Commitment to Reduce Hunger and Undernutrition in Developing Countries, *IDS Evidence Report 26*, Institute of Development Studies.

te Lintelo, D.J.H. and Lakshman, R. (2015) Equate and Conflate: Political Commitment to Hunger and Undernutrition Reduction in Five High-Burden Countries, *World Development*, 76, 280-292.

te Lintelo, D.J.H. (2019) *Developing Subnational Scorecards for Nutrition Accountability in Tanzania*, IDS Working Paper 530, Brighton: IDS

te Lintelo, D.J.H., Munslow, T., Pittore, K. and Lakshman, R. (2020) Process Tracing the Policy Impact of ‘Indicators’, European Journal of Development Research, 32, (4).

White, J. and Ypi, L. (2010.) “Rethinking the Modern Prince: Partisanship and the Democratic Ethos.” *Political Studies* 58 (4): 809–28.

Ypi, L. (2016) Political Commitment and the Value of Partisanship, *The American Political Science Review*, 110, (3), 601-613.

Chapter 5: Process tracing and values

5.1 Introduction

In a previous chapter, I explain the arguments of contextual empiricists and feminist philosophers of science who state that there is no use in lamenting the use of contextual values since science cannot be carried out without them. I define contextual values as values that belong to the social and cultural environment where science is done and include the norms, preferences, beliefs, or interests about what is acceptable or unacceptable practice, and which vary from community to community (Longino, 1990). The use of contextual values is inevitable whenever we engage in "ampliative inference". If the inference is ampliative—going beyond the facts—the hypothesis will be underdetermined by the facts, as the famous Duhem-Quine thesis teaches. Feminist philosophers use this argument to show that all kinds of values can and do enter science. But more importantly, scientific conclusions often reflect the perspectives and values of science practitioners (Longino, 1990).

What some feminist philosophers argue is that scientists express the contextual values inherent in their cultures and particular lives when they engage in a scientific activity (Longino, 2002: 135-136). One way to reduce the limitations this produces can be to try to ensure the values inherent in our assumptions are made explicit so that we better understand the kinds of roles that values are playing in scientific reasoning (Plutynski, 2018). In a previous chapter, I identify three ways contextual values enter in the production of and use of evidence: in the selection and collection of facts; the turning of facts into evidence; and the weaving of these together to confirm or disconfirm a hypothesis.

This chapter concerns the production and use of evidence for policy improvement via the tools of evaluation science. Usually, the aim is to gather facts to use as evidence to evaluate a claim that a policy or intervention worked or will work as intended. More generally, evaluation concerns the production of evidence to inform how a policy or intervention affects outcomes, whether these are intended or unintended (OECD-DAC, 2010).

My example is of a specific evaluation for the intervention ‘mHealth, Indonesia’ (World Vision International, 2015). mHealth pilots the use of mobile phones to improve the efficiency and quality of an existing nutrition service featuring community-based growth monitoring and nutrition counselling. mHealth more broadly encompasses a variety of interventions that aim to support governments to transform and modernise community health and nutrition services by introducing innovative mobile phone applications to support nutrition outcomes. The primary causal hypothesis of interest is that real-time data improves undernutrition monitoring. Current evaluation practice guidelines urge that this hypothesis be presented as a model that presents the step-by-step process by which the intervention is to produce its intended results, sometimes called a *theory of change*. In joint work I have done with others (Cartwright et al 2020) we recommend that a special kind of *process tracing* methodology be used to unpack finer-grained details of the causal process governing this black-box relationship between the putative cause – the availability of real-time data – and the desired outcome – better undernutrition monitoring.

My main concern is the use of this *process tracing* methodology to produce a theory of change and where it may fail. Theories of change are fallible representations of the real thing. They should not be confused for the actual causal process ‘out there’. The theories that are constructed, and the facts gathered and counted as evidence in favour of them, are affected by our perspectives and the values encoded in them. One concern is that we might be blinded to alternative kinds of evidence that support alternative theories of what might happen following a given intervention if we uncritically assume one perspective.

Given that theories are always underdetermined by whatever body of evidence is available[38], claiming that evidence supports a theory sufficiently to accept that theory without consideration of at least plausible alternatives from different perspectives is a failure at gathering and processing evidence properly. So other plausible characterisations of the process of interest should be described as accurately as possible.

---

[38] Underdetermined in the sense that the data available are always insufficient to determine what beliefs to hold in response to it (Longino, 1990.

What is interesting about the intervention I discuss is that despite the predicted success of using real-time data, the use of technology to make real-time data available is still relatively new, and it is inconclusive whether the use of this technology will lead to improvements in undernutrition in any specific setting. The potentials of real-time data are huge, but it is unwise to take unwarranted pride in technology since assumptions about its benefits to users are often not met. Value-driven decisions about what theories to consider and what evidence to bring to bear on refining the theory can blind us to alternative framings, including ones that capture a more comprehensive picture of the real-world workings of interventions and their consequences for those affected. What matters is that the overall evidence considers evidence for and *against* a hypothesis from different relevant perspectives. This requires a concerted effort to investigate all of the best available evidence about a preferred course of action.

This is an important topic if we are to better understand the specific role of values in the production and processing of evidence for policy improvement. Whether or not the production of evidence can and should be value free is a highly controversial claim involving many disputes. My intention is for the kind of detailed case study I offer to show how contextual values are hard to avoid and to suggest ways to watch out for where they enter, what roles they play, and what the implications are. I will do this by making *the kinds of roles* that values are playing more explicit in my case.

More broadly I believe my case defends the claim: value judgments are really unavoidable in the production of evidence for policy improvement. But this raises a problem. We don’t want to risk value judgments operating to drive our evidence to a predetermined conclusion. As a way to ameliorate this problem, following Plutynski (2018), Mertens and Wilson (2012) and others, I think that at the very least our value commitments ought to be made explicit so that they may remain open to revision in light of new evidence. This though is a big topic that I don’t go into in this dissertation.

If value judgments are not identified and appropriately made explicit, they have the potential to become significant barriers to the success of an intervention and its translation into important health benefits. I will outline a way to offer contextually appropriate support to assist the implementation of mHealth, making explicit some of the value judgments (e.g., ethical, political, and social) arising in the context of nutrition monitoring. This involves the development of a methodology to review and refine theories as we lay out in Cartwright et al (2020).

5.1.1 Chapter outline

First, I offer a detailed outline of the abstract form of a particular kind of causal-process-tracing theory. In a recent paper, 'Making predictions of programme success more reliable', Cartwright et al (2020), I, along with the other authors, set out the role of these causal-process-tracing theories (*pTocs*) for evidence gathering and assessment. I outline our approach to causal-process tracing theories and illustrate it with an example intervention, mHealth Indonesia. Second, I outline my concern with the use of a methodology that prioritises a single causal process. I aim to demonstrate as clearly as possible a number of places where values enter in the production of evidence and what implications this has. To do so, I offer illustrations of taken-for-granted assumptions about the use of technology and how it will interact with local factors that are likely to inform the production of evidence.

Besides explaining and defending our theory and illustrating it with the case study I developed, what I further want to do with this example is to correct a very important omission in our joint paper. We failed to bring to the fore an important requirement about constructing pToCs. That is the requirement that the total body of evidence for evaluating the claim that a pToC on offer will indeed describe the causal process that happens in a given setting if the related intervention is introduced should include not just evidence that directly confirms or disconfirms this claim but also evidence that rules out alternative theories of what will happen. In constructing our proffered example pToC for mHealth I did indeed consider alternative causal processes, but this is nowhere made explicit. Here I aim to make this requirement explicit and show how the inclusion of 'derailers' in the pToC goes some way to achieving this. As with all other aspects of the pToC, what possible derailers get noticed and included will be heavily influenced by the cultural values of the pToC designers and of the setting in which they work. What I intend to show is the need to engage with different value perspectives to identify priority issues (e.g., ethical, social, political) to support the refinement of theory.

5.2 Causal-process-tracing

One way to make headway on what counts as evidence for or against the hypothesis that a programme or intervention will achieve the targeted outcome in a targeted setting is to construct a programme theory. Programme theories refer to a structured set of assumptions about how an intervention is expected to work and how it is expected to influence processes of change. A causal-process-tracing theory (pToC) is a particular type, advocated by me and other authors in Cartwright et al (2020). It is a model that presents the step-by-step process by which an intervention is to produce its intended effects, including specific information about the “principles of change” for each step and the implications of these (Cartwright et al. 2020: 7). A pToC is useful both for evaluating post hoc whether an intervention produced its targeted effects and for predicting ex ante that it will do so. It is also more detailed than most applications of programme theory in the field of international development.

5.3 Middle-level theories

A causal process tracing approach often starts with a middle-level theory designed for the programme on offer. Middle-level pToCs help with applying evaluation findings from one setting to inform programme design in another. The underlying logic is to “go up” from the findings of how a programme worked in a specific setting to build a middle-level pToC that accommodates those findings and then “go down” to inform the local programme pToC for a new setting (Cartwright et al. 2020). This is what gives the approach the label middle-level theory: from an observed programme to middle-level theory and back down to a new programme theory. Following Cartwright at al,

> the purpose of a middle-level programme pToC is to present as much information as possible to help those making decisions about local use build a similar but much thickened pToC for the setting they are dealing with. Even

> where no general derailers are known, the programme pToC could have some empty boxes with question marks to remind these decision makers and their advisors to consider what these might be at each step in their setting (Cartwright et al. 2020: 30).

Applications such as mHealth involve the collection of data on nutrition. Such applications are considered as vital to the global effort to monitor undernutrition in infants and to respond to rapid changes in nutrition conditions (Barnett and Edwards, 2014: 3). Yet,

> far less is known about what happens to the data once they are generated. How can real-time data be aggregated, presented and disseminated most effectively and appropriately to trigger rapid responses and increase accountability and commitment to addressing undernutrition? What are the potential challenges and barriers to the effective use of real-time surveillance data? (Barnett and Edwards 2014: 3 in Munslow et al. 2016: 4)

We need to understand what factors are important in efforts to monitor undernutrition in infants. One important issue is that applications of mHealth may be prevented from working in different contexts. An application of mHealth will work differently depending on a variety of contextual factors, including social, political, ethical, and economic factors, for example, information on the contextual arrangements that need to be in place if the intervention is to achieve its intended effect. One central thing our approach achieves that is beyond usual applications of process tracing is that it includes specific information about the principles of change for each step and the implications of these (Cartwright et al. 2020: 7). This leads us to propose six types of information that go into a good theory of change if we want that theory to make clear what kinds of facts to gather as evidence about whether the intervention will produce the desired outcomes in a given setting. We outline these six types of assumptions in Cartwright et al (2020: 19):

1. The overall programme theory (as outlined above)
2. The middle-level principles that underwrite each step in the causal sequence
3. For each step, the support factors that must be in place for that step to help produce the next. Support factors are also called ‘interactive variables’ as their absence or presence determines whether there is an effect or moderates the scale of that effect.

If any support factor is missing along the way, and no substitute is in place, the whole process is endangered

4. For each step, the derailers that may prevent the next step or diminish the chances of it being successfully produced
5. For each step, what safeguards might protect the causal process from derailers
6. A general account of the range of application, i.e., the kinds of settings where the programme can be expected to work.

Let me explain more about each of these.

5.3.1 Assumption 1: Overall programme theory

An overall programme explains why a programme could initiate a process that can result in the desired outcome. This is what realist evaluation usually means when they call for the 'mechanism' of the programme to be described (Cartwright et al. 2020: 20). An overall programme theory may just be a verbal account of the leading ideas, or it may be an abstractly sketched theory of change or both.

For example,

> These theories may be simple in structure or they may be fairly elaborate. For an example of a relatively simple theory, consider mHealth programmes, where part of the theory refers to data accuracy. The theory is that technology will improve the accuracy of data on children's nutrition status and this in turn will promote a better response by key nutrition stakeholders, e.g. faster referrals to a midwife or other health services. This assumes that one of the reasons response rates have been historically low is the perceived inaccuracy of data on children's weights. We could call this the 'boy who called wolf' theory: if a signal has proven inaccurate in the past, we will stop using it (Cartwright et al. 2020: 20).

5.3.2 Assumption 2: Middle-level principles

Middle-level principles are sometimes called tendency principles. As we explain in our paper, these are

> principles that there is good reason to believe can hold across the range of settings in which the programme is expected to work. We clump principles of all

> these different degrees of generality together under the heading ‘middle-level theory’. Most will be tendency principles (Cartwright et al. 2020: 21).

Further,

> tendency principles often describe familiar psychological or sociological dispositions widespread across individuals or institutions, or widespread in individuals and institutions in specific settings. In this sense, tendency principles are not deterministic (Cartwright et al. 2020: 22).

Many middle-level principles are too general to play a helpful role in the middle-level programme theory. Think about the principle that a parent will tend to do what is in the best interest of a child. Other principles can be too specific. Think about the principle that technology is calibrated in such a way that produces accurate data. I go on to give more specific examples in my detailed case below.

5.3.3 Assumption 3: Support factors

Support factors are sometimes called moderators or interactive variables. Support factors help to produce the desired outcome. For example,

> [f]ew of the things we call ‘causes’ can produce their effects by themselves. They almost all need other factors to cooperate with them to make it likely that the cause will have the kind and/or size of result expected. We call these ‘necessary factors’ or ‘support factors’. (Others call them ‘helping factors’, ‘moderators’, or ‘interactive variables’) (Cartwright et al. 2020: 23).

A support factor may be unknown for different reasons. For example,

> At any stage the highlighted cause and its support factors are seldom the only set of causes that affect what happens next. Multiple other causes will be affecting the outcome independently of the ones implicated in the programme pathway, often pulling in different directions. The most we can expect from even a full set of mutually supporting factors is that they make the **contribution** expected at that stage: they pull the actual outcome in the right direction but do not dictate what the outcome will be (Cartwright et al. 2020: 26).

5.3.4 Assumption 4 and 5: Derailers

Derailers are things that can intervene in a programme to “stop a full set of causes from producing the expected contribution or substantially diminish the effects” (Cartwright et al. 2020: 28). Derailers are often unforeseeable and cannot be predicted in advance. But many are foreseeable. Program developers and implementers must identify what these derailers might look like so as to give the programme a better chance of succeeding (Cartwright et al. 2020: 29). Safeguards refer to the information that can be identified about how to prevent possible derailers and stop them from undermining the process.

5.3.5. Assumption 6: Range of application

By going through these six assumptions, we get more information about what the settings of a successful application will look like (Cartwright et al. 2020: 32). Adaptation of a pToC to suit the context in which it is intended can be an important step. The local conditions that applied in some other setting where the programme worked are unlikely to be the same as those that matter in a different setting. Without taking the local context into account, certain theories may not be supported by the local conditions (Cartwright et al. 2020: 39). Consider this point further:

> pToCs can be useful both for *ex ante* prediction and *post hoc* evaluation. In thickening a middle-level pToC for retrospective evaluation, it might be the case that an essential support factor is missing yet a positive effect is still observable. This would suggest that the middle-level pToC is not a good explanation of success for this setting, prompting the search for a better explanation (Cartwright et al. 2020: 40).

Middle-level pToCs are a key tool for using evaluation findings from one setting to inform programme implementation in another. The underlying logic is to ‘go up’ from the findings of how a programme worked in specific settings to build a middle-level pToC that accommodates those findings and then ‘go down’ to inform the local programme pToC for a new setting (Cartwright et al. 2020).

I demonstrate these assumptions in my example below. It is the mHealth example that I constructed for our joint paper. Unlike the work I have cited as joint, this example and the considerations surrounding it are my own original work. In my discussion, I will pay special attention to the importance of assumption 4: *derailers*. I have claimed that in gathering and processing evidence for a given hypothesis it is important to consider as accurately as possible alternative hypotheses. In this case this means plausible potential ways in which the intervention may fail to achieve the intended outcomes. This ensures that at least more than one characterisation of the process of interest is considered. As I said, I did indeed do this in constructing the pToC for mHealth offered in Cartwright et al 2020. But that is not noted explicitly. The considerations about alternative possible processes are buried in the derailers.

5.4 A new example: mHealth for nutrition monitoring in Indonesia

Here, I offer an example of mHealth Indonesia in support of a type of assessment that gathers evidence to test the single hypothesis of 'yes it did' / 'no it didn't' (Cartwright et al. 2020: 7). This approach is useful for bringing attention to the type of information that is needed if we are to make headway on the question of what kinds of facts counts as evidence for this kind of causal hypothesis (Cartwright et al. 2020: 7)[39].

My example is heavily based on a pilot application of mHealth for nutrition monitoring in Indonesia (Barnett et al. 2016). The pilot targeted a population of around 11,300 children under the age of 5 in North and East Jakarta, Surabaya, Pontianak and Sikka (World Vision International, 2016). mHealth was integrated into the existing national nutrition service delivery through the posyandu programme. Posyandus offer monthly service posts at sub-village level and are the lowest form of primary health care infrastructure in Indonesia (World Vision International, 2015). The posyandu emphasises community empowerment, and the organisation of posyandu activities are typically the responsibility of the village community. Support from heads of the community and community health committees are essential for the success of posyandu activities (Nazri et al. 2014). There is no financial requirement for the continued running of the posyandu (MoH, Indonesia, 2011). Funding comes from community members, donations, religious activities, social groups, private funding, commercial activities, and government sources (Berman, 1992). The posyandu was

---

[39] This example is also published in the paper, 'Cartwright, N. 2020. *Using middle-level theory to improve programme and evaluation design,* CEDIL Methods Brief. Oxford: CEDIL. The specific example of mHealth in Indonesia is my own work.

historically run by women who were part of the Women's Group for Empowerment and Family Welfare (known in Indonesia as the "PKK"). Historically, the position of working in a posyandu was highly respected within the community. The posyandu is now run by community health volunteers who are selected on the basis that they demonstrate a commitment to women's interests. So, the posyandu is run by the community for the community with technical support from the district (known as puskesmas)[40].

Interest in the collection of real-time data in Indonesia is huge. For example, Nokia corporation, a global leader in 5G, networks and phones, began funding a pilot in Indonesia in mid-2017 in partnership with UNICEF with the aim of transforming and modernising community health and nutrition services (NOKIA, 2021). The pilot introduces innovative mHealth technology to improve health and nutrition. In mid 2018, the ministry of Social Affairs worked alongside UNICEF to assist with the mapping and visualisation of vulnerable children in 27 urban centres, leveraging data collection technologies and platforms already established (NOKIA, 2021). In addition, World Vision recently partnered with Dimagi, a social impact company, and Grameen Foundation, a non-profit based in Washington DC, to pilot mHealth models with a view of scaling up (Dimagi, Inc, 2021).

Community health volunteers in the pilot study sites were provided with a mobile phone to replace a manual system. In 2013, World Vision Indonesia received technical support from the MOTECH Suite[41] to design a mobile phone-based application to address some of the challenges known to hinder nutrition service delivery in Indonesia: targeting key issues with growth monitoring and nutrition counselling in the community health clinic (World Vision International, 2015). Following Barnett et al. there are many issues with manual growth monitoring in Indonesia (Barnett et al. 2016). Some challenges include the incorrect categorisation of children's weight as normal, which leads to cases where children are missed from being offered nutrition support services. Other challenges are around the slow retrieval of children's details and subsequent delays in offering support services. There are

---

[40] I will now use the terms community health clinic and district health clinic, so the example is easily understandable to those without background reading.

[41] The MOTECH platform is an open-source enterprise software package that connects popular mHealth technologies to nutrition outcomes.

also challenges associated with the limited training and supervision of community health volunteers for manual growth monitoring status classification and uncertainty around how to calculate and interpret manual growth charts.

mHealth serves many purposes. Growth monitoring forms[42] for each child attending the community health clinic are completed by the community health volunteers and health facility workers and saved on the mobile phone application via general packet radio service[43]. The growth monitoring forms can then be accessed by health users in different agencies, such as the district health clinics (World Vision International, 2016). The mobile application processes growth and nutrition measurements automatically and flags the level of nutritional risk. It does so by classifying a child's nutrition and growth velocities through use of a standard z-score. In doing so, the application provides tailored nutrition messages that can be provided by community health volunteers to mothers in 'real-time', or at the time of the visit to the community health clinic (World Vision International, 2016). This process omits the role of community health workers manually working out the growth velocities and is thought (among other things) to improve the accuracy of the data.

The mobile application also serves other purposes, such as the generation of growth trend summaries, detailing areas of extreme prevalence of undernutrition or average weights of children under 5 in a particular location (World Vision International, 2015). Further, the application provides assessments of underlying illness and the current feeding practices of a child during home-based counselling sessions (World Vision International, 2015).

mHealth is considered to offer an innovative solution for community health volunteers and nutrition stakeholders to support better nutrition outcomes. Accordingly, between 2013 and 2015 a mixed-methods evaluation of mHealth was carried out by the Institute of Development Studies in partnership with World Vision Indonesia to better understand how mHealth was supporting these better nutrition outcomes. The evaluation and supporting documentation (Munslow et al. 2016) have been used to further unpack the context of

---

[42] This involves the collection of anthropometric measurements.

[43] GPRS

community-based monitoring in Indonesia, the current challenges, as well as how the mobile phone application was integrated into the posyandu.[44]

The evaluation offers an excellent insight into the conditions under which the mobile application brings about desired improvements in growth monitoring and nutrition counselling,[45] specifically, improvements in the 'accuracy, timeliness and responsiveness of growth monitoring', all of which are important for effective community-based growth monitoring (Barnett et al 2016: 13). The evaluation also provides a rich narrative around the context of nutrition services in Indonesia and motivations of various nutrition stakeholders (Barnett et al 2016: 18-19). 14 posyandus (in Sikka, North and East Jakarta) were selected to inform the evaluation (Barnett et al 2016: 32, 35). Qualitative and quantitative data are provided in the form of electronic databases, interviews and focus groups (Barnett et al 2016: 23-24).

Among the many important findings uncovered in Barnett et al's evaluation, 'mobile phone improved accuracy of growth monitoring status classifications', 'the mobile phone increased the timeliness of the growth monitoring in all posyandus', and '[o]f the mothers who received feedback, a significantly higher proportion (93 per cent) received it from [community health volunteers] with mobile phones' (Barnett et al 2016: 39). Some of the reasons for the mobile phone increasing feedback to mothers during the growth monitoring sessions were found to be context specific and related to issues such as 'trust in the feedback provided' and the 'objectivity of the calculations provided by the mobile phones' (Barnett et al 2016: 41). These are important findings and reflect just some of the contextual conditions under which the mobile application brings about desired improvements in growth monitoring and nutrition counselling (Barnett et al 2016: 18-19).

---

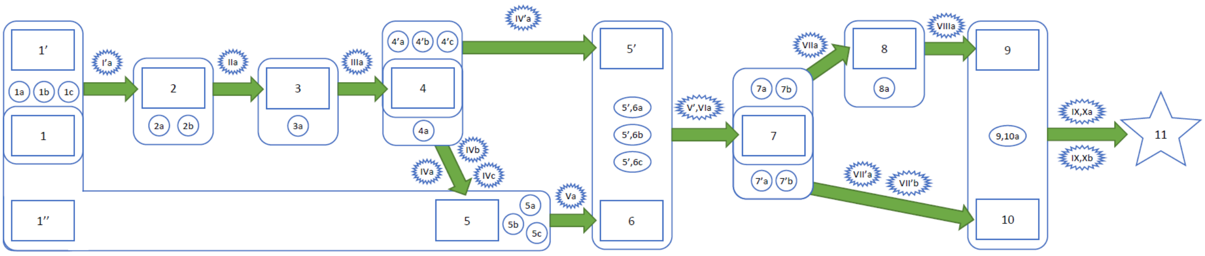
[44] A theory of change was constructed for Barnett et al's (2016) evaluation covering 14 posyandus in Indonesia, which has been extremely useful in reconstructing the six assumptions for a cpToC detailed in this report.

[45] This programme of work and evaluation was led by Dr Inka Barnett at the Institute of Development Studies, Brighton (see Barnett et al 2016: 18-19).

mHealth has the potential to help improve children's nutrition in other places where children can be weighed in community health clinics, but data is to be collected and curated and resources allocated at a higher (say, district) level.

Below is the causal process theory (pToC) that I constructed for Cartwright et al 2020, which follows our format for intervention assessment, using among other sources information from a series of follow-up visits with health stakeholders relevant in mHealth Indonesia (e.g., World Vision implementation staff, caregivers attending the clinics (posyandu), community health volunteers, government officials[46]). mHealth aimed for a number of outcomes: children brought to the community health clinics being better nourished; more accurate childhood nutrition data being available at district and higher levels; more efficient allocation of district resources; pressure on the government to address local problems of childhood nutrition. I have chosen for the illustration to focus on the outcome: 'children brought to the clinics are better nourished'. Similar causal theory (pToCs) can be constructed for the other outcomes as well. My mHealth casual theory (pToC) appears in Figure 1. The Figure and key are taken from our jointly authored paper.

Figure 1.



The key to Figure 1 is in Frame 1 below.

Cartwright et al (2020)

---

[46] This information can be found in the supplementary report Munslow et al 2016.

Frame 1,

Overall programme theory (Assumption Type 1): mHealth is administered and used by community health volunteers for growth monitoring in community health clinics where weight data is stored in mobile phone technology. The technology improves the accurate classification of infant growth status and promotes a timelier response from district health centres. This improves feedback and response to mothers attending the clinics and provides them with the advice and help they need. This leads to the outcome of 'children brought to the clinics are better nourished'.

Box 1. mHealth is administered

Box 1'. Community health volunteers are mandated to use it

Box 1''. District workers are mandated to monitor, curate and respond

Principle 1,1'-2: **Health workers tend do what they can in their clients' best interest**

Support factors:

1a. Community health volunteers have the capacity to use mHealth
1b. Community health volunteers agree that using mHealth is good for their clients
1c. Mothers and children attend the community health clinic on a regular basis

Derailers:

I'a. External pressure to not perform the task or other priorities prevail

Box 2: Infants' weight data is recorded in phones

Principle 2-3: mHealth technology does accurate calculations of growth status

Support factors:

2a. mHealth technology is well designed for the job
2b. Community health volunteers input the correct data in the correct format

Derailers:

IIa. Technology fails to operate

Box 3: Accurate classification of infants’ growth status by mHealth technology

Principle 3-4: Automated monitoring systems reduce error (e.g., there is a reduction in the misclassification of underweight infants with an automated monitoring system)

Support factors:

3a. Community health volunteers understand how to interpret results recorded by mHealth technology

Derailers:

III a. External pressure to not perform the task or other priorities prevail

Box 4: Accurate identification of underweight infants in the posyandu

Principle 4-5: Automated methods for doing so promote data submission

Support factors:

4a. Community health volunteers are able to submit data

Derailers:

IV a. External pressure to not perform the task or other priorities prevail

IV b. Other tools dominate and/or are seen as more useful

IV c. Technology fails

Box 5: More timely submission of data from local to district health clinics, especially in rural areas

Principle 1,1’,5-6: Health workers tend do what they can in their clients’ best interest

Support factors:

5a. District health staff know what to provide

5b. District health staff agree to the importance to clients of so doing

5c. District health staff are able to provide help

Derailers:

Va. External pressure to not perform the task or other priorities prevail

Box 6: District level help (referrals etc) provided to mothers where indicated by data

Principle 4-5’: Health workers tend do what they can in their clients’ best interest

Support factors:

4’a. District health staff understand the results

4’b. District health staff are able to provide feedback and advice

4’c. District health staff agree the importance of the feedback and advice for the clients

Derailers:

IV’a. External pressure not to do so or other priorities/goals

Box 5’: Timely feedback and advice to mothers in clinics

Principle 5’,6-7: People take in advice clearly given by people in positions of trust

Support factors:

5’,6a. Advice and help are clearly presented

5’,6b. Mothers can understand the advice

5’6c. Mothers trust community and district health institutions

Derailers:

V’,VIa. ?

Box 7: Mothers have knowledge and help they think can make their children better nourished

Principle 7-8 Mothers do what they think I in their children’s best interest

Support factors:

7a. Mothers think this is in children’s’ best interest overall

7b. Mothers can follow advice

Derailers:

VIIa. Conflicting norms or priorities

Principle 7-10: Mothers do what they think is in their childrens' best interest

Support factors:

7'a. Mothers think this is in children's best interest overall

7'b. Mothers and children can continue attending

Derailers:

VII'a. Conflicting norms or priorities

VII'b. Physical or economic preventions

Box 8: Mothers follow the advice

Principle 8-9: Children who are better nourished begin to grow more

Support factors:

8a. The advice and help are a good way to better nourish those children

Derailers:

VIIIa. Other factors (like illness) impede growth

Box 9: Undernourished children from clinic begin to be grow better

Box 10: Mothers and children continue attending clinic and getting advice/help

Principle 9,10-11: Children who are better nourished grow better

Support factors:

9,10a. The advice and help are a good way to better nourish those children across time

Derailers:

IX,Xa. Other factors (like illness) impede growth

IX,Xb. Failure at local or district level to supply advice/help

Box 11: 'Children brought to the clinics are better nourished'.

Cartwright et al (2020)

Let us now build this up one stage at a time. It is important to realise that this will not really involve a linear progression of thought. A general causal process theory (cptToC) is often a good example of 'grounded theory' where the theory emerges as it is developed. Each stage in the development informs the others. Work on one kind of assumption may provoke a rethinking of some assumptions made earlier, or as different assumptions are added, it may become clear that the initial sequence of intermediate stages needs to be modified. What matters is that at the finish, the cptToC as a whole is coherent and credible.

The simple input/output ToC looks like Figure 2.

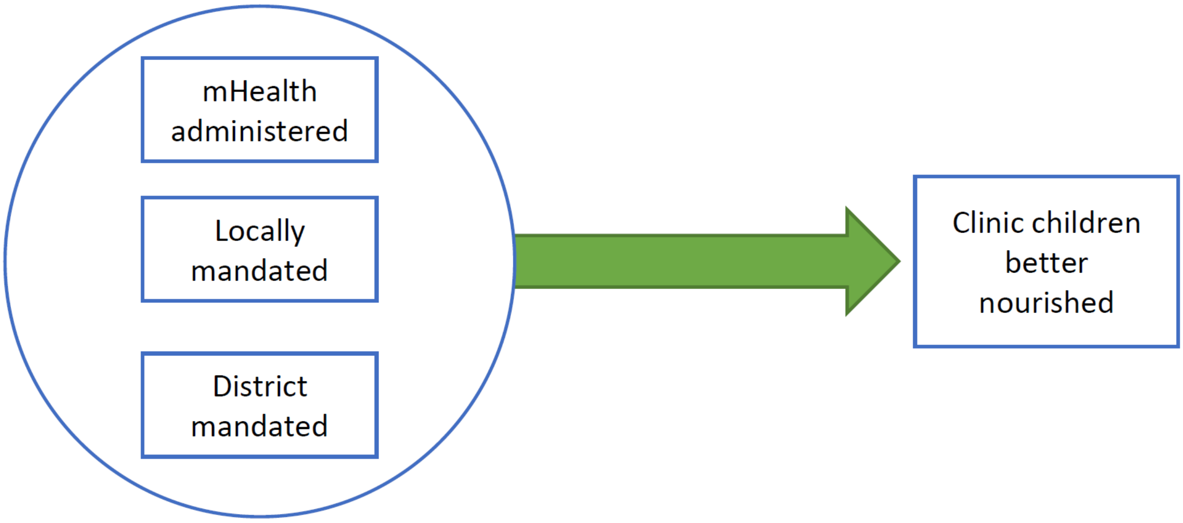


Figure 2.

Cartwright et al (2020)

Next, I introduce the significant causal steps in between, as in Figure 3. Doing the requisite thinking and research to figure out just what the process is supposed to be and how the programme input initiates it is a great part of the work in designing a programme that can be expected to carry though start to finish.

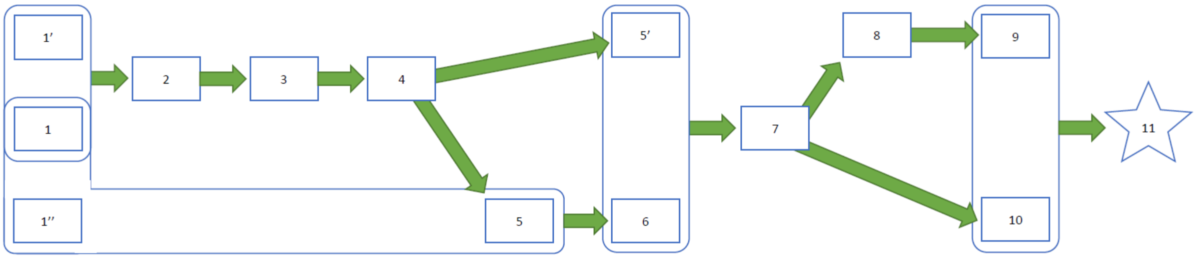


Figure 3.

Cartwright et al (2020)

Next, the support factors, derailers and safeguards must be added. Consider step 2-3 ‘Infants’ weight data is recorded in phones’ leads to ‘Accurate classification of infants’ growth status by mHealth technology’. This step will tend to work because of the principle that the ‘mHealth technology does accurate calculations of growth status’, as represented in Figure 4.

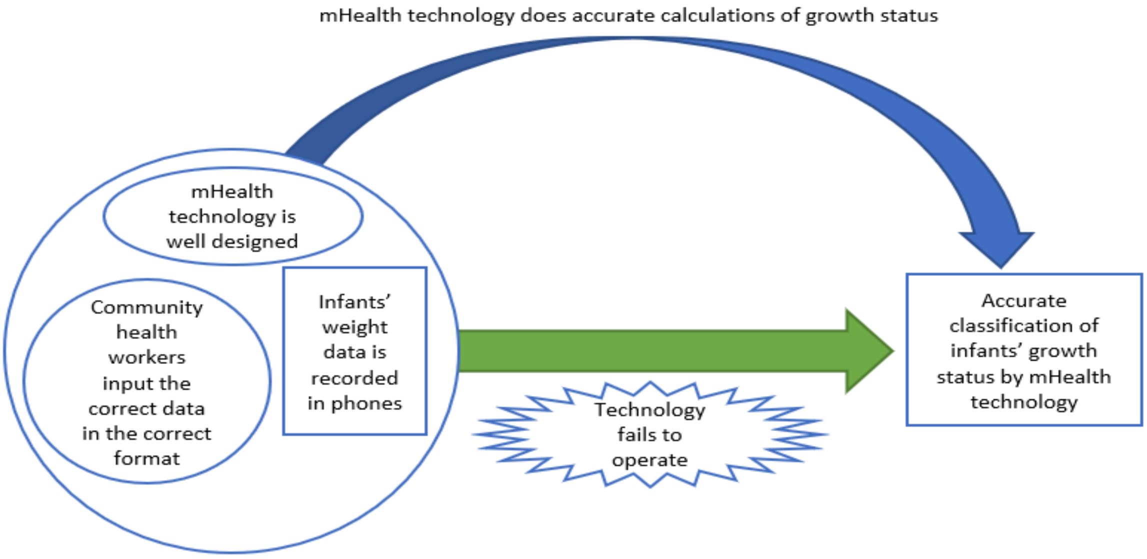


Figure 4.

Cartwright et al (2020)

However, to get from this step to the next the mHealth technology must be well designed for the job and community health volunteers must input the correct data in the correct format. The whole process will also be derailed if the technology fails to operate (represented in the jagged oval in Figure 4). For legibility, in the full general cptToC diagram for mHealth (Figure 1), all these descriptions have been abbreviated to their label as assigned in the key. Furthermore, the arrows representing the causal principles by which each step leads to the next have been omitted to keep the diagrams visually simpler, though these are detailed in the key. It is critical that principles are discovered and recorded, even if it may not be possible to include the arrows representing them in a legible diagram.

Building up the process in this way makes sense because of just how many things can go wrong or prevent the programme from achieving the targeted outcome. In the general cptToC for mHealth, there are over 21 factors to consider if the programme is to work as intended. For example, mothers must be able to understand the advice, mothers must think that this is in the children’s best interest overall, and mothers must trust the community and district health institutions, among other factors. Findings from Indonesia suggest for the middle-level pToC that mHealth programmes like this one are likely to work only if several key features are in place, such as health workers operating in the client‘s best interest and mothers‘ trust. Several derailers were found to undermine the programme there, including conflicting norms and priorities and external pressure not to perform tasks. In wider mHealth settings, programmes are only likely to work if the setting in which they are to be applied can find contextually relevant safeguards.

In the mHealth example, many of the support factors relate to the role of the community health worker and/or the district health staff: for example, 'community/district health staff have capacity‘ and ‘mothers understand the advice‘. The importance of these supporting factors raises evaluation questions related to the structure and functioning of key health services and the impact of new technology on an existing health system. If an evaluation were to discover that these essential features were missing, then this should be taken as evidence that the programme was not responsible for the outcome even if the hoped-for outcome did occur.

Beyond this, the mHealth middle-level pToC suggests several monitoring and evaluation indicators that may be adapted to future programme settings, for instance: correct application entry of weight into the mHealth application, timely entry of weight into mHealth application, delay in data sent from community health clinics to district health centres, mothers' feelings towards clinics and reattendance rates at community health clinics.

As always, decisions about how serious these problems might prove to be, whether they can be worked around and what are the costs and benefits and for whom of trying and failing (or trying something else or nothing at all) requires good sense and good judgment and can seldom be highly certain.

5.5 "Derailers" – evaluating alternative hypotheses

A derailer is any factor that may *prevent* the next step in the causal process represented in the theory or diminish the chances of it being successfully produced (Cartwright et al. 2020). It is necessary to describe any derailers as accurately as possible in order for us to be able to predict the potential ways in which a step may fail. This is an important step in our methodology. Evidence that possible derailers of the process described in the pToC on offer are (or will be or were) absent is evidence against the hypothesis that different processes happen other than the one described.

I know about the particular derailers listed in Figure 1 because I drew on different user perspectives to inform our theory (Munslow et al. 2016). This involved the conduct of stakeholder interviews with programme and evaluation specialists at World Vision Indonesia (Munslow et al. 2016). One thing that preliminary research suggests is that different stakeholders put emphasis on different things. Mothers interviewed about the intervention emphasise the trustworthiness of the information provided by the technology, and whether it provides them with the information they need. Mothers reported feeling anxious about attending the monthly weigh-in sessions. They also reported that the process of meeting targets can be an intimidating process (Munslow et al. 2016). District health staff put importance on whether the data will overburden them with what is an already information heavy system. Community health volunteers focus on the ease of inputting the data and whether it will improve communication with the district hospitals.

Logically, the absence of derailers can be thought of as the presence of support factors. In the joint paper we defended classifying the two differently because thinking in terms of what can go wrong, what can derail the process at any step, directs attention to further factors you might miss if you only think in terms of what's needed to give the cause at that step the power to produce the outcome at the next step. But there's a deeper reason to focus on derailers. Doing so is an indirect way to introduce consideration of alternative hypotheses: what might occur to set a different process off at this stage? This is the reason I am highlighting derailers here. The articulation of a derailer allows us to consider the evidence likely to push *against* our preferred process by providing evidence that a different process occurs. As I said above, it is always best to rule out alternative hypotheses to the extent possible. Considering direct evidence for or against just a single process—the one under test—is not a good thing. And looking at what might happen from different perspectives is a crucial step in coming to recognize what other processes might occur instead of the one modelled in the pToC under test.

One limitation of a causal-process-tracing theory is that it offers a single theory for an intervention. A single causal process for an intervention is considered in the narrow context of asking a programme implementer to present a model of what is *intended* to happen (Cartwright et al. 2020). This emphasises one perspective (e.g., the technology provider) of how an intervention is proposed to work—thus prioritising the implementers perspective in providing this model. What this does not stress is that we should not simply rely on the model as laid out by those implementing the programme as our guide for looking for evidence that the interventions will achieve the intended outcomes.

Unsurprisingly, intervention realities are often quite different to those we envision (Cartwright et al. 2020). Practical constraints of time, data, financial resources, staff incentives and available expertise are important factors here[47]. There is much valuable information that comes from the expertise of implementers. But there is no one perspective

---

[47] Jos Vaessen makes this point.

guaranteed to achieve a good body of evidence. A causal-process theory is a *fallible* model. It needs to stand up to scrutiny from different perspectives.

If we are interested in predicting the likely success of an intervention, we need to be concerned with how success is seen from different perspectives. It is well known that if an intervention doesn't align with the cultural and political values of a particular setting then it will likely fail. There are a whole host of factors that need to be in place if something will work – exploring what these are is an important part of building a good causal process theory.

When it comes to informing what counts as evidence, we don't want to only look for evidence from one perspective. One risk is that we may overemphasise some facts as a result of our entrenchment in a single perspective. We want to gather evidence on a broad range of factors that are important to different users within the system. We also don't want to assume the benefits apply to all users, which is important to making an intervention better. For example, the technology could be modified to better suit the purposes of the end-users. However, I do not address this point here.

In the following section, I will first outline the basic principles of process tracing. I will second outline how I came up with the derailers in Figure 1 and demonstrate their usefulness in predicting success (or failure).

My examples fulfil an illustrative purpose only. It may not be that the real-life cases would look exactly like the ones I present. Or perhaps better alternative hypotheses could be imagined. What is vital is that those tasked with developing a causal process theory actively seek potential alternatives. This involves a consideration of different value perspectives on what is important, and to whom (Munslow et al. 2016). Hence, I am suggesting we avoid committing to only *one* value perspective.

5.6 Process tracing logic ala Bayesianism: its basic principles

Many put Bayesianism at the methodological foundation of process tracing. This allows inferences[48] to be made about a single case by assessing alternative hypotheses or theories in light of evidence, which is what I have been advocating. Bayesianism is concerned with how facts we find can help us change, or up-date, our confidence in a hypothesis. It tells us that to improve confidence in a hypothesis seek both pieces of evidence in support of or against the hypothesis under consideration and pieces of evidence against alternatives. In our case this implies identifying what evidence is expected to be found for each part of the causal process theory under test (e.g., directly confirming or disconfirming evidence) and for plausible alternative theories (e.g., disconfirming evidence) (Beach and Pedersen 2013: 95). Ideally, we would like every piece of evidence to point in the same way, but this is rarely the case in social science where “we must often make inferences based on the accumulated weight of evidence from many clues, none of which is strongly decisive” (Fairfield and Charman, 2017: 6).

Bayes formula is this: $P(e/h) = P(h/e)\ P(h) \mid P(e) = P(e/h)\ P(h) \mid [P(e/h)\ P(h) + P(e/-h)\ P(-h)]$
For the Bayesian logic of updating then we need the following information:

1. The extent to which a hypothesis predicts the evidence offered in its support, $P(e/h)$
2. The likelihood of the hypothesis prior to the evidence, $P(h)$
3. The likelihood of that evidence given the alternative hypotheses, $P(e/-h)$

What evidence we need is based on two evaluation questions. *Certainty* asks, what do we have to find if the hypothesis is true? The degree of certainty of empirical predictions depends on the expected facts that are likely to be found in the real world (e), when the hypothesised part of the mechanism actually exists (h). We surely do want to find that the facts that with high certainty should obtain if our hypothesis is true do obtain. So, we want to establish whether or not e is true for those facts e for which $p(e/h)$ (Van Evera, 1997: 30).

---

[48] In process tracing case studies, inferences are made by assessing the match between predicted and actually found evidence for each part of the mechanism (Beach, 2016: 468). We update our confidence in the primary explanation on the basis of the evidence. The question addressed at the formalisation stage of a causal mechanism is how much does our initial confidence in the existence of each part of this mechanism increase (or decrease) when bringing out new evidence? (Befani and Mayne 2014, 23).

Uniqueness asks, is there a plausible alternative for finding the facts observed? The uniqueness of empirical predictions depends on the expected observations that are likely to be found in the real world (e), when the hypothesis is false (¬h). Uniqueness is represented as the probability of *false positives* p(e/¬h) (Bennett, 2015: 276-278).

Based on the degrees of certainty and uniqueness of empirical observations, four kinds of tests can be conducted (Van Evera, 1997). The four tests are:

- **Straw-in-the-wind tests**: Failure or passage of this test neither lends strong support for or against the hypothesis (*low certainty and low uniqueness*)
- **Hoop tests**: Failure to pass a hoop test can be disqualifying for a hypothesis but passing the hoop test does not necessarily lend strong support for the hypothesis (*high certainty and low uniqueness*)
- **Smoking-gun tests**: Passing a smoking gun test lends strong support for the hypothesis, whereas failure does not necessarily lend strong support against the hypothesis (*low certainty and high uniqueness*)
- **Double decisive tests**: Passing a double decisive test lends strong support for the hypothesis while also lending strong support **against alternative hypotheses** (*high certainty and high uniqueness*).

Straw-in-the-wind tests are the weakest and offer very little support for a hypothesis (Collier, 2011: 826). *A smoking-gun test* can strongly support a hypothesis, but failure to pass it does not reject it. “If the hypothesis passes, it also substantially weakens rival hypotheses” (Collier, 2011: 827). *A doubly decisive test* offers the most support for a hypothesis, since they “provide strong inferential leverage that confirms one hypothesis and eliminates *all* others” (Collier, 2017: 827: emphasis added). For example, a hypothesis (or theory) is more probable if the facts it predicts to hold do hold and if they could not have been predicted from the best rival hypothesis available (Elman and Elman, 2002 in Rholfing, 2014).

In addition, we need to figure out for the case at hand what specific alternatives to h need to be taken seriously, as Fairfield and Charman (2017) explain in a recent paper on the use of process tracing methodology in social science:

> Whereas most treatments compare a single working hypothesis *H* against it(s) logical negation, *~H,* **we advocate identifying one or more concrete rival hypotheses.** This approach is critical in social science, because *~H* generally will not be a well-defined proposition—*H* could fail to hold in an essentially infinite number of ways. The more specific the hypothesis is, the more possibilities are embodied in *~H.* Directly assessing likelihoods of the form *P*(*E*/*~H I*) will be practically impossible if we have not first contemplated what concrete possibilities *~H* might actually entail (Fairfield and Charman, 2017: 4, emphasis added).

Ideally, we want to find facts that obtain *if* the hypothesis is true. But we don't want those same facts to also support the alternative hypothesis. We need facts that the alternative hypothesis *cannot* equally account for. It is too hard to figure out what *would* be the case if not h (-h). So, we need specific hypotheses to tell us what specific facts to expect if the hypothesis is false: -h is too open ended.

5.7 Predicting success: steps in the causal process

Illustration 1: step 2-3 'Infants' weight data is recorded in phones' leads to 'Accurate classification of infants' growth status by mHealth technology'

Consider step 2-3 in Figure 1. If this step is likely to hold, we would predict to see evidence of the following support factors being present: (1) mHealth technology is well designed for the job; (2) Community health volunteers input the correct data in the correct format; (3) Community health volunteers understand how to interpret results recorded by mHealth technology. We would also predict the absence of the following derailers: (4) Technology fails to operate; (5) External pressure to not perform the task or other priorities prevail.

We would expect to find evidence of 1-3 and so we may pass the hoop test, as these observations have *high certainty* since the hypothesised step 2-3 predicts all three observations. So, if we observe that these three obtain, we have evidence that provides a degree of confirmation for the claim that ‘Infants’ weight data is recorded in phones’ leads to ‘Accurate classification of infants’ growth status by mHealth technology’.

High uniqueness requires consideration of any plausible alternative for finding these specific observations. An alternative hypothesis could be set out as follows: Infant’s weight data is recorded in phones, but this does not lead to an accurate classification of infants’ growth status because the technology fails to operate. We need to actively look for evidence of ‘the technology failing to operate’. If we find evidence of the technology failing to operate then it offers support for the alternative hypothesis and weakens the likelihood of step 2-3.

The technology could fail to operate properly in at least two ways. One is that it simply fails to produce an output—no classification is provided. A second is that a classification is provided but this is not accurate, either because the technology does not calculate according to the formula prescribed or because that formula does not produce accurate classifications of growth status. With respect to the first, we know that technology often fails to operate in real cases in the field, for a great variety of reasons. These include inadequate internet, intermittent electricity supply, faulty equipment, improper training, among others.

As to the second, the formula that is supposed to be used employs a z-score to classify weight. But background knowledge about nutrition studies tells us that there are reasons to be critical of the World Health Organisation endorsed z-score as a classification of weight in infants (see Svedberg, 2000). Anthropometric data is a specific type of information relating to the scientific study of the measurements and proportions of the human body. Anthropometric data is created specifically by tracking the weight-for-age and weight-for-height of individuals attending the clinics, as recommended by the World Health Organisation. The software utilised by most mHealth applications expresses the distance between an individual infants’ weight and the average weight of the comparable infants in the reference population by a standard z-score. A z-score can show a negative value which indicates an infant is below the mean. Or a positive value which indicates an infant is above the mean. So, the use of a z-score means that the infant can be labelled as moderate or severe. There are lots of advantages to data being accumulated in this way. The amount of data that can be stored is huge, continuous and there is always up-to-date information. However, there are several problems with the use of a z-score.

One problem is that additional information about the mother or infant is sometimes required to make a full assessment of infant vulnerability (e.g., lack of iron in the mother and/or infants). This type of information is not picked up by use of a standard z-score[49]. The z-score only offers limited information on whether the type of malnutrition is acute, moderate, or severe, with no additional information on the context surrounding the undernourished person. The data produced via z-score may not prioritise other types of information perhaps previously collected at the posyandu, such as other features that surround the likelihood of malnourishment (e.g., mother's wellbeing).

In order to take the evidence to support step 2-3, we have to assume that anthropometric data offers a valuable source of knowledge about the growth status of the child. If we have reason to doubt the classification offered by a z-score for the context it is intended to be used in, then the emphasis we put on the evidence informing the step is reduced. Recognising the alternative hypothesis is important since we may not automatically question the validity of the classification scheme of the phone if we are committed to the use of standard z-score to measure and monitor undernutrition. If we take for granted the assumption that the z-score offers a valuable classification scheme, then it closes off the evidence that we produce in consideration of the hypothesis.

If we uncritically accept the assumption that anthropometric data is a valuable source of information for understanding infant health, then we wouldn't be naturally inclined to look for evidence to test this assumption[50]. The obvious next place to start is to gather more evidence starting with the facts of the case. Why is a z-score particularly useful for classifying the growth status of these children coming to this posyandu? What is it about a z-score that makes it particularly fit for purpose?

---

[49] It would only be picked up if the infant's weight faltered below the critical threshold for weight as per the classification system used, then the infant may be treated at the district hospital where further issues may be picked up.

[50] Anderson (2004) claims that value-judgments *ought* to be open to revision in light of the evidence. That is, we can hold our values open to revision in light of new evidence.

What matters is the overall evidence, considering both evidence for and against and from differing perspectives. This requires a concerted effort to investigate all of the best available evidence *against* a preferred course of action. It is in this way that using different value perspectives expands the evidence base.

Illustration 2: step 7-10 ‘Mothers have knowledge and help that they think can make their children better nourished’ leads to ‘Mothers and children continue attending clinic and getting advice/help’

Consider step 7-10 in Figure 1, ‘Mothers have knowledge and help that they think can make their children better nourished’ leads to ‘Mothers and children continue attending clinics and getting advice/help’. If this step is likely to hold, we would predict to see evidence of the following support factors being present: (1) Mothers think this is in children’s best interest overall; (2) Mothers and children can continue attending. We would predict to see an absence of the following derailers: (3) Conflicting norms and priorities; (4) Physical or economic preventions; (5) Failure at local or district level to supply further advice/help. Note that (3) and (4) are in the original, (5) I add now.

An alternative hypothesis could be set out as follows: ‘Mothers do *not* continue to get help/advice because of a failure at the district level to supply it’. We need to actively look for evidence of ‘a failure to supply help/advice at the district level’. If we find evidence of this failing, then it offers support for the alternative hypothesis and weakens the likelihood of step 7-10.

We know that this scenario is likely because our background information about the case tells us that the community health volunteers are concerned about the level of support that they receive from the district hospitals. The quality of growth monitoring services provided at a community health clinic (e.g., posyandu) is heavily dependent on the midwife from a district health clinic (e.g., puskesmas) being present for a posyandu session (Mize, 2012). If she is not able to attend, then women who need replenishment of supplies or an injection will be without protection (Mize, 2012). If the midwife does not attend, then this is a huge problem as the mothers can’t get the support they need. So even if the mothers *do* attend the clinics on a regular basis, they may not continue to get the help and advice they need if the district midwife does not also attend.

This compels us to look for additional evidence. For example, evidence that focusses on different aspects of the context, such as the role of the district midwife. One place to start is to investigate what it is about the community healthcare volunteers' use of mHealth that makes it more likely that the district midwife attends the posyandu. There may be completely different reasons why the midwife does not attend on a monthly basis that the use of technology will not solve. Given our background assumptions we look for different evidence relevant to the hypothesis under scrutiny. If we take for granted that the introduction of a new technology will automatically incentivise the district hospitals to use the data, then we may not find other evidence relevant to the step under scrutiny, such as the constraints that district midwives experience.

Another example of an alternative hypothesis is as follows[51]: ‘Mothers do have knowledge and help they think can make their children better nourished, but this does not lead to mothers and children continuing to attend clinic and getting advice/help because although mothers think what they have received so far can make their children better nourished, they don’t trust the changes to the system of growth monitoring or are wary of the mHealth technology being used to interpret the progress being made in their infants’ growth. We need to actively look for evidence of ‘mistrust in or wariness towards the changes to the system of growth monitoring’. If we find evidence of this failing, then it offers support for the alternative hypothesis and weakens the likelihood of step 7-10.

We know that this scenario is likely because our background information about the case tells us that attendance at the weigh in is driven by mothers who honestly want to know, “How is my child doing?” and are compelled to attend on this basis (Munslow et al. 2016). Mistrust translates into experience of stress or uncertainty, which, in turn, affects the likelihood of

---

[51] The point is that we need a specific hypothesis to tell us what specific facts to expect if the hypothesis is false: -h is too open ended.

attendance at the posyandu. Or the likelihood that mothers will invite the healthcare volunteers to their homes if necessary. Something then needs to be done to reduce the mistrust. If something is not done, then attendance may continue to falter putting the infant at further risk.

This compels us to look for additional evidence. For example, evidence that focuses on the motivations of the mother in attending the posyandu. One place to start is to investigate how the mothers feel about the use of the new technology. Are they satisfied? Do they have any concerns/stresses? What motivates mothers to continue attending?

Illustration 3: step 5’,6-7 ‘District level help (referrals etc) provided to mothers where indicated by data’ and ‘Timely feedback and advice to mothers in clinics’ leads to ‘Mothers have knowledge and help they think can make their children better nourished’

Finally, consider step 5’,6-7 in Figure 1, ‘District level help (referrals etc) provided to mothers where indicated by data’ leads to ‘Timely feedback and advice to mothers in clinics’ leads to ‘Mothers have knowledge and help they think can make their children better nourished’**.** If this step is likely to hold, we would predict to see evidence of the following support factors: (1) Advice and help are clearly presented; (2) Mothers can understand the advice; (3) Mother’s trust community and district health institutions. We would predict to see an absence of the following derailer: (4) external pressure by competing monitoring systems. Note that (1-3) are in the original, (4) I add now.

An alternative hypothesis could be set out as follows: Mothers do *not* have knowledge and help they think can make their children better nourished because of external pressure by competing monitoring systems at the district level. We need to actively look for evidence of ‘competing monitoring systems’. If we find evidence of such systems, then it offers support for the alternative hypothesis and weakens the likelihood of step 5’,6-7.

It is assumed that the district level will be able to provide help as long as there is reliable data to tell them where to allocate their time and resources. But background knowledge about the case tells us that there are reasons to be cautious of implementing new systems into already burdened systems. If we take for granted the system already in place into which the new technology is being implemented, then we may not understand to look for evidence of competing monitoring systems. Background knowledge about the case tells us that staff members in the district hospitals (including district midwives) experience huge amounts of stress due to the level of data that they are required to already process (Munslow et al. 2016; Berman, 1992).

We are required to look for additional evidence. For example, evidence that focuses on the existing monitoring systems in place at the district level. Are there additional systems in place that are still currently in use? What are these monitoring systems? It stands to reason that the mHealth technology may fail to produce the intended step if there are competing monitoring systems in place. So, this is important evidence that weighs against the preferred course of action as stated in the causal theory. And evidence that we need to look for.

5.8 Support factors and values

Some support factors are easier to identify from different points of view. For example, if you are a programme developer used to working in a particular setting where you assume that volunteers are trained in the use of technology then you may not see the relevance of training as a support factor. Whereas, if you are a volunteer working in a posyandu then you may understand that effective training is a necessary factor for the success of a programme. We need different value perspectives to inform our understanding of the factors that need to be in place in different contexts to ensure the success of an intervention. The identification of significant factors (e.g., cultural, social, ethical) will be best done by including different moral and cultural perspectives.

Impartiality and "the inevitability thesis"

At this point, I will clarify the relationship between Douglas' views (as outlined in detail in chapter 2), and the relationship between impartiality and the inevitability thesis. There appears to be a tension between saying values inevitably play a role and urging a method that appears designed to protect the assessment of evidence from undue influence by values.

It is my contention that evidence is *not* value-free. Here, I argue that the production and use of evidence via an evaluation to inform social policy is value laden. Following Heather Douglas, I argue in chapter three that the evidence available to us at a given time may be insufficient to determine what beliefs we should hold in response to it. I now want to further argue that the production of evidence to test an intervention and its likely success is also value laden. This is because there is uncertainty left by the evidence that leaves room for values to influence whether we claim an intervention is likely to be successful. For example, if we inherently believe that nutrition monitoring via technology is something valuable in itself, then we may be more likely to conclude that the intervention is successful given the presence of evidence to support that claim, and accept the remaining uncertainty left by the evidence.

Here, I contemplate quite a complicated problem: how to factor lessons about the roles of values in science into process tracing techniques for assessing the likelihood that policy interventions will work in a given setting. The problem tackled is whether an intervention is going to have the intended effect when implemented, and/or why it will not have that effect. Traditionally, process tracing is a method that appears designed to protect the assessment of evidence from undue influence of values, which seems in keeping with the value-free ideal. However, I want to argue that we need to critically engage with the intervention - the complex technical, social, cultural, and political values underpinning an intervention – and interrogate the values of the intervention itself to inform the evaluation.

This is a different way of working. There is no clear separation between the intervention and the evaluation. The evaluator cannot maintain a detached viewpoint or complete impartiality from the goals of the intervention. The evaluator needs to interrogate the values underpinning an intervention and produce facts to test a hypothesis from different relevant value perspectives. Given the uncertainty left by the evidence it may never be possible to claim with a reasonable degree of certainty that an intervention is successful, however the evaluator needs to gather sufficient facts/data to test the hypothesis as rigorously as possible. Producing sound and plausible alternative hypotheses (taking account of the different relevant value perspectives) will likely inform this process. A key point is the importance of determining up front (e.g., prior to the onset of the evaluation) the level of certainty that is required by the evaluation commissioner. This is a key point that I have not addressed in my chapter, and one that offers some clarity on the relationship between Douglas' views and the relationship between impartiality and inevitability.

Briefly, the evaluation must engage with values directly (as per Douglas' distinction) to understand whether there is a sufficient weight of evidence to accept the hypothesis under scrutiny (e.g., the primary causal hypothesis). We cannot answer whether the evidence base is sufficient without first understanding what level of certainty is required. If the evaluation finds an intervention to be successful (when it is in fact not), to what extent will this have an unintended effect on the recipients of that intervention? If it is likely that the undue acceptance of an intervention will have a negative impact on recipients' health or wellbeing, for example, then we must demand a higher degree of certainty, as per Douglas' argument. These are value laden issues that cannot be avoided. Whether we set the threshold higher (or lower) depends on what (or who) we care about, what mistakes we want to avoid, and how much time/resources we have to work with.

What this tells us about the relationship between impartiality and inevitability is as follows: an evaluator (or the evaluation more broadly) utilises social science methods like causal process tracing which are typically taken to require the evaluator to occupy a detached or impartial view on the nature of the claim to be evaluated. While causal process tracing might appear to be a method that is designed to protect the assessment of evidence from undue influence by values, I have argued that there is a need to engage with values directly. There is no obvious answer to important (and value laden) questions on which the method relies: how much evidence to gather, what facts are important, what is significant, whose interests' matter, and to what degree, and who gets to decide. Thus, one reasonable way forward seems to involve the critical assessment of such questions through a consideration of different relevant value perspectives.

5.9 Conclusion

In gathering and assessing evidence about a hypothesis it is important to look for and consider not only evidence that speaks directly for or against the hypothesis but also evidence that rules out the plausible alternatives. Different perspectives reflecting different values will throw up different alternative hypotheses. If we approach an evaluation from only one perspective, we are likely to miss considering alternative hypotheses that really need to be taken into account. Given that hypotheses are always underdetermined, claiming that the evidence supports the hypothesis without consideration of at least plausible alternatives from different perspectives is a failure to process evidence properly.

I have demonstrated as clearly as possible a number of places where *value perspectives* enter in the production of evidence and what implications this has. If we are interested in predicting the likely success of an intervention, then we shouldn't avoid ways to better consider the chances of the intervention succeeding and look at it from different value perspectives. For instance, a technology company and a community may focus on very different factors and make very different assumptions about what things are like. We should find a way to consider these different perspectives in our causal process theories and not blindly commit to one over another.

This raises a problem though. A method like process tracing has to find a way to deal with conflicts of values and not unintentionally lend support to one set of values over another. In response to this problem many have argued that at the very least, our value commitments ought to be made explicit so that they may remain open to revision in light of new evidence (Plutynski, 2018; Mertens and Wilson, 2012). I endorse this view, but I do not undertake defending it in this dissertation. Rather I aim to make clear one way that different perspectives, and the values that come with them, actually ought to enter into the process of gathering and evaluating evidence. A detailed look at how to deal with conflicting perspectives is beyond what I can undertake here.

References

Anderson, E. (1995b), 'Knowledge, human interests, and objectivity in feminist epistemology', *Philosophical Topics,* **2**3, 27–58.

Anderson, E. (2002) Situated Knowledge in the Interplay of Value Judgments in Scientific inquiry, in P. Gardenfors, J. Wolenski and K. Kijania-Plackek, eds., In the Scope of Logic, Methodology and Philosophy of Science, vol. 2, 497-517.

Anderson, E. (2004) Use of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce, *Hypatia*, 19, (1).

Barnett, I., Sulistyo, S. S., Befani, B., KariSari, K., Sharmin, S. and Dewi, D. (2016). *Mixed-Method Impact Evaluation of a Mobile Phone Application for Nutrition Monitoring in Indonesia*. Institute of Development Studies.

Barnett, I. and Edwards, D. (2014) Barnett, I. and Edwards, D. (2014) *Mobile Phones for Real-time Nutrition Surveillance: Approaches, Challenges and Opportunities for Data Presentation and Dissemination*, IDS Evidence Report 75, Brighton: IDS

Beach, D. and Pedersen, R. (2013). *Process Tracing Methods: Foundations and Guidelines*: University of Michigan Press.

Bennett, A. (2015). Appendix: Disciplining our conjectures, systemizing pro- cess tracing with bayesian analysis. Process Tracing: From Metaphor to Analytic Tool, 276-298. Cambridge: Cambridge University Press.

Berman P. (1992) Community-based health programmes in Indonesia: the challenge of supporting national expansion. In: Frankel S, ed. *The Community Health Worker: Effective Programmes for Developing Countries*. England: Oxford University Press; 62-87.

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Williams, R.B. (2020), *Making predictions of programme success more reliable,* CEDIL Methods Working Paper. Oxford: CEDIL.

Cartwright, N. (2020). *Using middle-level theory to improve programme and evaluation design,* CEDIL Methods Brief. Oxford: CEDIL

Collier, D. (2011) Understanding Process Tracing, *PS: Political Science and Politics*, 4, 823-830.

Crasnow, S. (2017) Process tracing in political science: What's the story, *Studies in History and Philosophy of Science,* 62, 6-13.

Dimagi, Inc (2021) *Case study, Deploying MOTECH Suite for MNCH and Nutrition programs in 10+ countrie*s. Available at: https://www.dimagi.com/case-studies/mhealth-world-vision-motech/ (Accessed: 20 October 2019).

Elman, Colin, Elman, Miriam Fendius. (2003). "Liberal International Relations Theory: A Scientific Assessment." Pp. 159-204 in Progress in International Relations Theory: Appraising the Field, edited by Elman, C., Elman, M. F. Cambridge, MA: MIT Press.

Fairfield, T. and Charman, A. (2017) Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats, *Political Analysis*, LSE eprints, 1-19.

Longino, H. (2002) *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Longino, H. (1990) *Science as Social Knowledge*. Princeton, NJ: Princeton University Press.

Mertens, D. M., and Wilson, A. T. (2012). *Program evaluation theory and practice: A comprehensive guide.* The Guilford Press.

Ministry of Health Indonesia. *(2011) Pedoman Umum Pengelolaan Posyandu (Posyandu General Guide)*. Jakarta, Indonesia.

Munslow, T., Barnett, I. and Dewi, D. 2016. The Potential of M-health for Improved Data Use: Workshop Report, IDS Evidence Report 201. Brighton: Institute of Development Studies.

Mize, L. (2012) The Paradox od Posyandu: Challenges for Community Health Services in Indonesia; Literature Review. AusAID HRF.

Nazri, C., Yamazaki, C., Kameo, S., Herawati, D., Sekarwana, N., Raksanagara, A. and Koyama, H. (2016) Factors influencing mother's participation in Posyandu for improving nutritional status of children under-five in Aceh Utara district Aceh province, Indonesia, *BMC Public Health.*

NOKIA, (2021) *Nokia and Unicef collaborate to boost healthcare to 58 million children in Indonesia.* Available at: https://www.nokia.com/about-us/sustainability/nokia-and-unicef-collaborate-to-boost-healthcare-to-58-million-children-in-indonesia/ (Accessed: 28th June 2021).

OECD-DAC, (2010) *DAC Guidelines and Reference Series, Quality Standards for Development Evaluation*. Available at: https://www.oecd.org/development/evaluation/qualitystandards.pdf (Accessed 20th October 2019).

Plutynski, A. 2018. *Explaining Cancer: Finding Order in Disorder*, Oxford University Press: Oxford.

Rholfing, I. 2014 Comparative Hypothesis Testing Via Process Tracing, *Sociological Methods and Research*, 43, (4), 606-642.

Russett, B. (1993). *Grasping the democratic Peace: Principles for a post-cold war world.* Princeton, NJ: Princeton University Press.

Svedberg, P. (2000) *Poverty and Undernutrition: Theory, Measurement and Policy*, Oxford University Press: Oxford.

Van Evera, S. (1997) *Guide to Methods for Students of Political Science,* New York: Cornell University Press.

World Vision International. 2018. *Global mHealth Report*. Available at: https://www.wvi.org/sites/default/files/2018%20mHealth%20Report-FINAL-electronic.pdf (Accessed: 20 October 2019).

World Vision International. 2016. *mHealth: Indonesia*. Available at: https://www.wvi.org/sites/default/files/Indonesia%20mHealth%20Factsheet%202016.pdf (Accessed: 20 October 2019).

World Vision International. 2015. *mHealth: Indonesia*. Available at: https://www.wvi.org/sites/default/files/Indonesia%20mHealth%202014%20v5%20pdf.pdf (Accessed: 20 October 2019).

Chapter 6: Conclusion

I have argued that value judgments are often unavoidable in the collection and processing of evidence. To ignore values is to make fact collecting antiseptic: it prevents the collection of facts or new ideas to test a particular claim. Several conclusions can be drawn from my research to improve the production and use of evidence, especially for policy improvement.

One concern is that we should not assume the legitimacy of one value perspective over another. Offering an account of the facts from a particular value perspective may restrict how we understand an important topic, like health or undernutrition, where there is a lot at stake. The challenge is not that the account on offer is untrue, but that it is incomplete. This has the implication of making one set of claims the *only* story. I argue that what matters is that the total body of evidence for evaluating a claim considers evidence for and *against* a claim from different relevant value perspectives. This requires a concerted effort to investigate all the best available evidence about a preferred course of action: including evidence that goes against our core suppositions. We should not be limited by unexamined assumptions since this has the effect of narrowing what we see. The specific role for values in the collection and processing of evidence that I have focussed on most is as follows: values are mediated by our background assumptions and inform what facts are relevant to our investigations, what facts get used as evidence, and what facts we emphasise.

We must be conscious that the kind of conceptualisation we bring to our investigations may be limited (Longino, 1992: 334). What is more, the kind of conceptualisation that we bring to our investigations can influence the evidence we obtain in support of a claim (Anderson, 2002). For example, nutrition surveillance interventions may cast technology as a solution to weak and underperforming health systems. Or, in emergency settings such as the outbreak of Ebola in 2014, medical interventions may endorse an assertive response to the viral spread of infection. I have argued that these are loaded with value-laden assumptions about the best course of action to tackle the issue at hand. While an intervention can be appropriate for the problem it is designed to address, we must remain critical of the fact that it is inevitably framed by *particular* interests and values. There's no reason to assume that a particular point of view is the right one, or the only relevant one, even if it is widely accepted. We need to increase our willingness to see beyond the status quo and actively seek alternative perspectives of what will happen. Since the tendency to overlook diverse value perspectives can undermine efforts to intervene successfully.

Perhaps more of a concern is that, if a value perspective becomes the accepted way of thinking, we may lose sight of plausible alternatives of what will happen. A value perspective may achieve dominance through replication, advocacy or the silencing or dismissal of other viewpoints. This, however well-meaning, directly benefits certain interests while potentially excluding or worse, harming others. For example, we may assume a default position about a new technology given the enticing benefits it could bring, and inadvertently end up prioritising evidence that supports this value-laden assumption. This may have the adverse effect of marginalising different viewpoints in the policy process or removing the ability of certain stakeholders to engage at all, depending on their openness and understanding of technology. As I have demonstrated, value-laden assumptions held by a particular individual or organisation can lead to the misrepresentation, whether under or over representing certain facts, or indeed the dismissal of facts that may count as evidence with a different set of assumptions. Developing ways to avoid a single value perspective in the collection and processing of evidence for use in policy is a key area for further study.

At the very least our value-laden assumptions ought to be made explicit so that they may remain open to revision in light of new evidence. Value-laden assumptions can be held open to revision and examination by considering *other* normative assumptions, perspectives, and worldviews in the collection and processing of evidence. In the examples in my thesis, I discuss the inclusion of different value schemes in the processing of evidence for use in nutrition policy. One suggestion is to include different values and understandings brought to the study of health and nutrition by major stakeholders: patients, their caregivers, clinical staff, civil society groups, umbrella organisations, businesspeople, policy makers, people from the media, and so on. Different value schemes inform where else we should look and what else is relevant in terms of properly testing a set of claims. We mustn't only draw on the substantive knowledge of major stakeholders to evaluate how effective an intervention is likely to be in achieving a particular outcome. In addition, we must draw on the knowledge of major stakeholders to actively challenge what effectiveness might look like from different critical perspectives. It is important that we leave ourselves open to surprise: we do not want to ‘rig the game’ in favour of our preferred outcomes (Anderson, 1995: 51).

I have already spent a lot of time detailing the inevitability of value-laden assumptions and so will not return to them in any detail here. It is uncontentious that values inform policy: what is problematic is that certain values may be given priority above others. What I will focus on instead is a way of thinking about values in production and use of evidence, especially for policy improvement. There is much more work to be done to identify, question, and justify the assumptions and values that ultimately influence policy and have implications for different values and interests. The issue is *what perspectives are relevant and how to combine them*. This means consulting all the relevant perspectives feasible and in the right ways. This immediately raises the question: *which are the relevant perspectives, are they being used in an appropriate way, and how should we decide*.

Longino famously argues that what matters is the diversity of the contextual values represented. She carefully describes the characteristics of ideal epistemic communities, claiming that it is easier to identify when values inappropriately influence scientific reasoning when those values are different to one’s own (Longino, 1990). Longino argues for the inclusion of diverse values and different perspectives that provide a sort of check against biasing values and interests to ensure that they do not inappropriately influence scientific reasoning (Longino, 1990). Beyond just arguing for a variety of points of view, feminist standpoint theorists argue for the special importance of including minority researchers and the subjects of research in research teams since they can have special epistemic access (Intemann, 2005). For feminist standpoint theorists, it is the specific experience of minorities and not just the diversity of social positions that is epistemically beneficial (Intemann, 2005). For instance, a person’s particular experience is what is more likely to identify the limitations with background assumptions that have been unnoticed. Standpoint theorists argue for the inclusion of minority researchers and subjects of research since they have special epistemic access. This is different to the claim that the subjects of research should reflect the diversity of our culture and conditions, considering race, ethnicity, gender, age, and so on. I do not take a stand either way; and much more empirical research would be necessary to do so.

Consider my example of the Ebola outbreak in West Africa in 2014, as discussed in chapter three. There is something about both the experience of anthropologists *and the diversity* of their position that played a fundamental role in understanding the spread and containment of Ebola. The anthropologists had professional experience in medical response yet *also* experience in the religious aspects of multicultural societies. The substantive knowledge of anthropologists was useful to point out different facts relevant to the understanding of the topic. This included facts on people, on lifestyle, on burials, on spiritual disposition, on reincarnation, and on the afterlife. It was the anthropologist's experience of medical settings, but also of West African society, that showed limitations with background assumptions that had previously gone unnoticed. Without the involvement of anthropologists, scientists may not have gathered facts on diverse aspects of the social context that were instrumental in understanding the containment of Ebola.

Further research on this topic will have to think about both epistemic and moral issues. I have argued that it is background assumptions that make facts evidence or not. Sometimes a person’s background assumptions are blatantly false by all reasonable standards. So, their implications about what should be taken into consideration in the full body of evidence are disallowed. Sometimes a person’s background assumptions may be true but morally inadmissible—like, the best way epistemically to get causal conclusions is by lying to people about what treatment they are being given in a randomised controlled trial. This background assumption may be true but there are moral reasons for not allowing it to guide us in how to gather evidence. Consider the infamous Tuskegee Study of Untreated Syphilis, as discussed in chapter two. There are serious ethical implications because the researchers consciously withheld information on possible treatment from research participants. Even long after a treatment option was available. Yet, the study was pursued on the assumption that this was the best way to get causal conclusions.

Consider the pilot intervention of ‘mHealth Indonesia’ once more. The cultural perspective of mothers was critical to understand whether the intervention is likely to succeed. Mothers offered important background information about the cultural and political setting that was useful to test the underlying assumptions of the intervention. Where the perspective of mothers was consulted, further lines of investigation were explored that may have otherwise been unattainable. Including a cultural perspective is epistemically beneficial as it provides the investigation additional evidence to test a claim. We surely want to include cultural perspectives so that we increase the likelihood of as many people as possible benefiting from intervention. However, perhaps there are situations where we should refrain from seeking a value perspective. For example, there are different kinds of mothers. Some are likely to be more vulnerable than others. If it is our belief that vulnerable mothers have special epistemic access, then we would surely want to target those mothers. However, perhaps we should hesitate at doing so if there can be predictable ways that those mothers may be harmed.

The value perspective of mothers makes it more likely that an intervention will succeed; however, this is not the sole purpose for inviting involvement. I argue that we want to involve a diversity of relevant perspectives for the purpose of considering *plausible alternatives of what will happen*. We don't want to overstate the importance of an intervention. If we really value the collecting of alternatives, then we must find ways to imagine what they might be.

We need evidence that rules out alternative theories of what will likely happen so that we don't end up overstating the likelihood of the effect. We also need evidence that rules out alternative theories to understand the factors working against the intervention. However, in doing so, we must find ways to remain critical of the assumptions underpinning the intervention: what makes something the *right* thing to do, and for whom? (Pawson and Tilly, 1997)

Failing to consult all the relevant perspectives can have epistemic and moral consequences for how likely the intervention is to be successful. If we want to understand whether we are intervening successfully then there are two important questions. There is an epistemic question: are the claims underpinning the programme well supported by the evidence? And there is a moral question: is the action of intervening the right thing to do? (Sen, 1981) We need both types of assessment to appropriately evaluate whether we ought to intervene in something like undernutrition. These have long been recognised as important questions in international development (Sen, 1981, Gasper, 2012; Sumner, 2007). Yet the relevance of these questions remains a key direction for further research into values in development. I hope to be part of future research on this topic.

References

Anderson, E. (1995) Feminist Epistemology: An Interpretation and a Defence, *Hypatia*, 50-84.

Anderson, E. (2002) Situated Knowledge in the Interplay of Value Judgments in Scientific inquiry, in P. Gardenfors, J. Wolenski and K. Kijania-Plackek, eds., In the Scope of Logic, Methodology and Philosophy of Science, vol. 2, 497-517.

Gasper, D. (2012) *The Ethics of Development*, Edinburgh University Press Books: Edinburgh.

Intemann, K. (2005) Feminism, underdetermination, and values in science, *Philosophy of Science*, 72: 1001-12.

Longino, H. (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton, NJ: Princeton University Press.

Pawson, R. and Tilly, N. (1997) *Realistic Evaluation*, London: Sage.

Sen, A. (1981) ‘Accounts, Actions and Values: Objectivity in Social Science’, in C. Lloyd (ed.), Social Theory and Political Practice: Wolfson College Lectures 1981. Oxford: Clarendon Press, 87-107.

Sumner, A. (2007) What are the Ethics of Development Studies? *Institute of Development Studies Bulletin*, 38, (2), 59-68.

Bibliography

Anderson, E. (1995) Feminist Epistemology: An Interpretation and a Defence, *Hypatia*, 50-84.

Anderson, E. (2002) Situated Knowledge in the Interplay of Value Judgments in Scientific inquiry, in P. Gardenfors, J. Wolenski and K. Kijania-Plackek, eds., In the Scope of Logic, Methodology and Philosophy of Science, vol. 2, 497-517.

Anderson, E. (2004) Use of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce, *Hypatia*, 19, (1).

Anoko, J. (2014). “Communication with Rebellious Communities during an Outbreak of Ebola Virus Disease in Guinea: An Anthropological Approach.” http://www.ebola-anthropology.net (Accessed 29th June 2021)

Atkinson, A. B. (2001) The Strange Disappearance of Welfare Economics, *Kyklos*, 54, (2/3), 193-206.

Barnett, I., Sulistyo, S. S., Befani, B., KariSari, K., Sharmin, S. and Dewi, D. (2016). *Mixed-Method Impact Evaluation of a Mobile Phone Application for Nutrition Monitoring in Indonesia*. Institute of Development Studies.

Barnett, I. and Edwards, D. (2014) Barnett, I. and Edwards, D. (2014) *Mobile Phones for Real-time Nutrition Surveillance: Approaches, Challenges and Opportunities for Data Presentation and Dissemination*, IDS Evidence Report 75, Brighton: IDS

Beach, D. and Pedersen, R. (2013). *Process Tracing Methods: Foundations and Guidelines*: University of Michigan Press.

Bennett, A. (2015). Appendix: Disciplining our conjectures, systemizing pro- cess tracing with bayesian analysis. Process Tracing: From Metaphor to Analytic Tool, 276-298. Cambridge: Cambridge University Press.

Berman P. (1992) Community-based health programmes in Indonesia: the challenge of supporting national expansion. In: Frankel S, ed. *The Community Health Worker: Effective Programmes for Developing Countries*. England: Oxford University Press; 62-87.

Biddle, J. (2013) State of the field: Transient underdetermination and values in science, *Studies in History and Philosophy of Science*, 44, 124-133.

Bradburn, Cartwright, N. and Fuller (2016) A theory of measurement, Working Paper, Centre for Humanities Engaging Social Science and Society (CHESS), Durham.

Cartwright, N. (2020). *Using middle-level theory to improve programme and evaluation design,* CEDIL Methods Brief. Oxford: CEDIL

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Williams, R.B. (2020), *Making predictions of programme success more reliable,* CEDIL Methods Working Paper. Oxford: CEDIL.

Cartwright, N. and Runhardt, R. (2014) Measurement, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 31-47.

Cartwright, N. (2008) *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Chopra, D. (2014) Towards Gender Equality with Care-Sensitive Social Protection, *IDS Policy Briefing 49,* Brighton: IDS.

Churchman, C. W. (1948) Statistics, Pragmatics, and Induction, *Philosophy of Science*, 15: 249-268.

Clements, M. and Hendry, D. (1999) On winning forecasting competitions in economics. Spanish Economic Review (1), 123-160.

Collier, D. (2011) Understanding Process Tracing, *PS: Political Science and Politics*, 4, 823-830.

Crasnow, S. (2017) Process tracing in political science: What's the story, *Studies in History and Philosophy of Science,* 62, 6-13.

Crasnow, S. (2014) Feminist Standpoint Theory, Philosophy of Social Science: A New Introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford.

Daston, L. and Gallison, P. (2007) *Objectivity*, Zone books: London.

Deaton, A. and Cartwright, N. (2018) Understanding and misunderstanding randomised controlled trials, *Social Science and Medicine*, 210, 2-21.

Deaton, A. and Paxson (1998) Aging and Inequality in Income and Health, American Economic Review: Papers and Proceeding, 88, (2), 288-253.

Decancq, K. and Lugo, M. A. (2010) Weights in Multidimensional Indices of Wellbeing: An Overview, Leuven: Katholieke Universiteit Leuven/Oxford: Dept. of Economics, University of Oxford, 30 August 2010.

Department of Health (1995) *Child Protection: Messages from Research*, London: HMSO.

Di Bucchianico, Maria Elena (2009) *Modelling high temperature superconductivity: A philosophical inquiry in theory, experiment and dissent.* PhD thesis, London School of Economics and Political Science: United Kingdom.

Dimagi, Inc (2021) *Case study, Deploying MOTECH Suite for MNCH and Nutrition programs in 10+ countrie*s. Available at: https://www.dimagi.com/case-studies/mhealth-world-vision-motech/ (Accessed: 20 October 2019).

Douglas, H. (2000) Inductive Risk and Values in Science, *Philosophy of Science*, 67, (4), 559-579.

Douglas, H. (2008) The Role of Values in Expert Reasoning, *Public Affairs Quarterly*, Vol. 22, No. 1, pp. 1-18.

Douglas, H. (2009) *Science, Policy, and the Value-Free Ideal*, University of Pittsburgh Press: Pittsburgh.

Douglas, H. (2014) *Values in Social Science*, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 162-185.

Duhem, P. (1956) Aim and Structure of Physical Theories, (first published 1914) trs. Phil- ip Weiner. Princeton. N: Princeton University Press.

Efstathiou, S. and Mirmalek, Z. (2014) Interdisciplinarity in Action, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 233-248.

Elliot, K. (2013) Douglas on values: From indirect roles to multiple goals, Studies in History and Philosophy of Science, 44, 375-383.

Elman, Colin, Elman, Miriam Fendius. (2003). "Liberal International Relations Theory: A Scientific Assessment." Pp. 159-204 in Progress in International Relations Theory: Appraising the Field, edited by Elman, C., Elman, M. F. Cambridge, MA: MIT Press.

Fairfield, T. and Charman, A. (2017) Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats, *Political Analysis*, LSE eprints, 1-19.

Fairhead, J. (2016) Understanding Social Resistance to the Ebola Response in the Forest Region of the Republic of Guinea: An Anthropological Perspective, African Studies, Review, 59, (3).

Gasper, D. (2012) *The Ethics of Development*, Edinburgh University Press Books: Edinburgh.

Gere, C. (2017) *Pain, Pleasure and the Greater Good: From the Panopticon to the Skinner Box and Beyond*, The University of Chicago Press: Chicago and London.

Gigerenzer, G. and Marewski, J. N. (2014) Surrogate Science: The Idol of a Universal Method for Scientific Inference, Journal of Management, https://doi.org/10.1177/0149206314547522

Gillespie, S., Haddad, L., Mannar, V., Menon, P., & Nisbett, N. (2013). The politics of reducing malnutrition: Building commitment and accelerating impact. The Lancet, 382(9891), 552–569. http:// dx.doi.org/10.1016/S0140-6736(13)60842-9.

Gouldner, L. H. and Williams, R. C. (2012). The choice of discount rate for climate change policy evaluation, Climate Change Economics, 3, (4), 12 pages.

Hammergren, L. (1998). Political will, constituency building, and public support in rule of law programs (No. PN-ACD-023). Center for Democracy and Governance, USAID.

Hempel, C. G. (1965) Science and Human Values, *in Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* New York: The Free Press, 81-96.

Hunger and Nutrition Index (2021), *Hunger and Nutrition Commitment Index*, Institute of Development Studies. Avaiable at: http://www.hancindex.org (Accessed: 28th June 2021).

IFPRI. (2014). Global nutrition report 2014: Actions and accountability to accelerate the world's progress on nutrition. Washington: International Food Policy Research Institute.

Intemann, K. (2005) Feminism, underdetermination, and values in science, *Philosophy of Science*, 72: 1001-12.

Intemann, K. and de Merlo-Martin, I. (2010) Social values and scientific evidence: the case of the HPV vaccines, *Biology and philosophy*, 25, (2), 203-213.

Jeffrey, R. C. (1956) Valuation and Acceptance of Scientific Hypotheses, *Philosophy of Science*, 22: 237-246.

Kabeer, N. (2012) Women's Economic Empowerment and Inclusive Growth: Labour Markets and Enterprise Development, SIG Working Paper, 2012/1, International Development Research Centre/Development for International Development.

Khosrowi, D. (2019) Trade-oofs between epistemic and moral values in evidence-based policy, *Economics and Philosophy*, 35, (1), 49-78.

Larner, W., and R. Le Heron. 2004. Global Benchmarking: Participating 'at a Distance' in the Globaliz- ing Economy. In *Global Governmentality: Governing International Spaces*, ed. W. Larner and W. Walters, 212–232. London: Routledge.

Levi, I. (1962) On the Seriousness of Mistakes, *Philosophy of Science*, 29: 47-65.

Longino, H. (2014) Individuals or Populations? Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 102-123.

Longino, H. (2002) *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Longino, H. (1992) Taking Gender Seriously in Philosophy of Science, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, (2), 333-340: The University of Chicago Press on behalf of the Philosophy of Science Association.

Longino, H. (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton, NJ: Princeton University Press.

Longino, H. (1979) Evidence and hypothesis: An analysis of evidential relations, *Philosophy of Science,* 46, (1), 35-56.

Maxmen, A. (2015) How the Fight Against Ebola Tested a Culture's Traditions, National Geographic, https://www.nationalgeographic.com/adventure/article/150130-ebola-virus-outbreak-epidemic-sierra-leone-funerals-1 (Accessed 29th June 2021).

McMullin, E. (1983) Values in Science, in Peter D. Asquith and Thomas Nickles (ed.), *Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association, Volume 1.* East Lansing: Philosophy of Science Association, 3-28.

Merry, S.E. 2011. Measuring the World Indicators, Human Rights, and Global Governance. *Current Anthropology* 52: 83–95.

Mertens, D. M., and Wilson, A. T. (2012). *Program evaluation theory and practice: A comprehensive guide.* The Guilford Press.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harpercollins.

Ministry of Health Indonesia. *(2011) Pedoman Umum Pengelolaan Posyandu (Posyandu General Guide)*. Jakarta, Indonesia.

Munro, E. (1999) Protecting children in an anxious society, *Health, Risk & Society*, 1:1, 117-127, DOI: 10.1080/13698579908407011

Munro, E. (2005) Improving practice: child protection as a systems problem, *Children and Youth Services Review*, 27, (4), 375-391.

Munslow, T., Barnett, I. and Dewi, D. 2016. The Potential of M-health for Improved Data Use: Workshop Report, IDS Evidence Report 201. Brighton: Institute of Development Studies.

Mize, L. (2012) The Paradox od Posyandu: Challenges for Community Health Services in Indonesia; Literature Review. AusAID HRF.

Myers, J. (1994) *The Backlash: Child Protection under Fire,* London: Sage.

Nazri, C., Yamazaki, C., Kameo, S., Herawati, D., Sekarwana, N., Raksanagara, A. and Koyama, H. (2016) Factors influencing mother's participation in Posyandu for improving nutritional status of children under-five in Aceh Utara district Aceh province, Indonesia, *BMC Public Health.*

NOKIA, (2021) *Nokia and Unicef collaborate to boost healthcare to 58 million children in Indonesia.* Available at: https://www.nokia.com/about-us/sustainability/nokia-and-unicef-collaborate-to-boost-healthcare-to-58-million-children-in-indonesia/ (Accessed: 28th June 2021).

OECD-DAC, (2010) *DAC Guidelines and Reference Series, Quality Standards for Development Evaluation*. Available at: https://www.oecd.org/development/evaluation/qualitystandards.pdf (Accessed 20th October 2019).

Parker, W. (2014) Climate Change, Philosophy of Social Science: A new introduction, Cartwright, N. and Montuschi, E., Oxford University Press, Oxford pp 31-47.

Parton, N., Thorpe, D. & Wattam, C. (1997) *Child Protection: Risk and the Moral Order*, London: Macmillan.

Pawson, R. and Tilly, N. (1997) *Realistic Evaluation*, London: Sage.

Putnam, H. (1981) *Reason, Truth and History*, Cambridge University Press: Cambridge.

Quine, W. V. O., 1951, "Two Dogmas of Empiricism", Reprinted in *From a Logical Point of View*, 2nd Ed., Cambridge, MA: Harvard University Press, pp. 20–46.

Plutynski, A. 2018. *Explaining Cancer: Finding Order in Disorder*, Oxford University Press: Oxford.

Reiss, J. (2008) Methodology and expert judgment in evidence-based economics: lessons from the CPI controversy, Error in Economics, 1st Edition, London: Routledge.

Reiss, J. (2013) *Philosophy of Economics: a contemporary introduction*, London: Routledge.

Reiss, J. (2016) *Error in Economics: Towards a More Evidence-Based Methodology*, London: Routledge.

Reiss, J. (2017) Fact-value entanglement in positive economics, Journal of Economic Methodology, 24, (2), 134-149. DOI: 10.1080/1350178X.2017.1309749.

Rholfing, I. 2014 Comparative Hypothesis Testing Via Process Tracing, *Sociological Methods and Research*, 43, (4), 606-642.

Rudner, R. (1953) The Scientist Qua Scientist Makes Value Judgments, *Philosophy of Science*, 20: 1-6.

Russett, B. (1993). *Grasping the democratic Peace: Principles for a post-cold war world.* Princeton, NJ: Princeton University Press.

Sandmo, A. (2017) Tony Atkinson 1944-2017: A lifetime commitment to the study of inequality, The European Journal of the History of Economic Thought, 24: 3, 612-623.

Sen, A. (1981) 'Accounts, Actions and Values: Objectivity in Social Science', in C. Lloyd (ed.), Social Theory and Political Practice: Wolfson College Lectures 1981. Oxford: Clarendon Press, 87-107.

Stanko, E. A. (1997) Safety Talk: Conceptualising Women's Risk as a "Technology of the Soul", *Theoretical Criminology*, 1: 479-499.

Stewart, A. J., Copeland, A. P., Chester, N. L., Malley, J. E., & Barenbaum, N. B. (1997). *Separating together: How divorce transforms families.* Guilford Press.

Stern, N. (2006). *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

Stiglitz, J. (2009) The anatomy of murder: Who Killed America's economy? Critical Review, 21, 329-339.

Sumner, A. (2007) What are the Ethics of Development Studies? *Institute of Development Studies Bulletin*, 38, (2), 59-68.

SUN Movement (2015) *SUN Civil Society Network*. Available at: https://scalingupnutrition.org/sun-supporters/sun-civil-society-network/ (Accessed 29th June 2021)

Svedberg, P. (2000) *Poverty and Undernutrition: Theory, Measurement and Policy*, Oxford University Press: Oxford.

te Lintelo, D.J.H., Munslow, T., Pittore, K. and Lakshman, R. (2020) Process Tracing the Policy Impact of 'Indicators', European Journal of Development Research, 32, (4).

te Lintelo, D.J.H. (2019) *Developing Subnational Scorecards for Nutrition Accountability in Tanzania*, IDS Working Paper 530, Brighton: IDS

te Lintelo, D.J.H. and Lakshman, R. (2015) Equate and Conflate: Political Commitment to Hunger and Undernutrition Reduction in Five High-Burden Countries, *World Development*, 76, 280-292.

te Lintelo, D.J.H. Haddad, L., Lakshman, R. and Gatellier, K. (2013) The Hunger And Nutrition Commitment Index (HANCI 2012) Measuring the Political Commitment to Reduce Hunger and Undernutrition in Developing Countries, *IDS Evidence Report 26*, Institute of Development Studies.

Van Evera, S. (1997) *Guide to Methods for Students of Political Science,* New York: Cornell University Press.

Waring, Marilyn. 1990. If women counted. San Francisco: HarperCollins.

White, J. and Ypi, L. (2010.) "Rethinking the Modern Prince: Partisanship and the Democratic Ethos." *Political Studies* 58 (4): 809–28.

Wilkinson and Leach (2015) Briefing: Ebola Myths, Realities, and Structural Violence, African Affairs, 114, (454), 136-148.

World Health Organisation (2014) *Ebola Response Roadmap Situation Report*. Available at: http://apps.who.int/iris/bitstream/handle/10665/137424/roadmapsitrep_31Oct2014_eng.pdf;jsessionid=954615B086633BFBD94675D4433374D4?sequence=1 (Accessed April 1, 2021).

World Health Organisation (2015) *Surveillance strategy during Phase 3 of the Ebola response, Emergency Guidance.* Available at: http://apps.who.int/iris/bitstream/handle/10665/192997/the World Health Organisation _EVD_Guidance_Sur_15.1_eng.pdf?sequence=1 (Accessed April 1, 2021).

World Health Organisation (2015b) *Factors that contributed to undetected spread of the Ebola virus and impeded rapid containment*. Available at: https://www.who.int/news-room/spotlight/one-year-into-the-ebola-epidemic/factors-that-contributed-to-undetected-spread-of-the-ebola-virus-and-impeded-rapid-containment (Accessed April 1, 2021).

World Health Organisation (2016) *An R&D Blueprint for action to prevent epidemics: Plan of Action, May 2016*. Available at:

https://reliefweb.int/report/world/rd-blueprint-action-prevent-epidemics-funding-coordination-models-preparedness-and (Accessed June 28, 2021).

World Health Organisation (2017) *How to conduct safe and dignified burial of a patient who has died from suspected or confirmed Ebola or Marburg virus disease*. Available at: https://www.who.int/publications/i/item/the World Health Organisation-EVD-Guidance-Burials-14.2 (Accessed April 1, 2021).

World Health Organization (2021). *Ebola Virus Disease*. Available at: https://www.who.int/health-topics/ebola#tab=tab_1 (Accessed April 1, 2021).

World Vision International. 2018. *Global mHealth Report*. Available at: https://www.wvi.org/sites/default/files/2018%20mHealth%20Report-FINAL-electronic.pdf (Accessed: 20 October 2019).

World Vision International. 2016. *mHealth: Indonesia*. Available at: https://www.wvi.org/sites/default/files/Indonesia%20mHealth%20Factsheet%202016.pdf (Accessed: 20 October 2019).

World Vision International. 2015. *mHealth: Indonesia*. Available at: https://www.wvi.org/sites/default/files/Indonesia%20mHealth%202014%20v5%20pdf.pdf (Accessed: 20 October 2019).

Ypi, L. (2016) Political Commitment and the Value of Partisanship, *The American Political Science Review*, 110, (3), 601-613.

Zahle, J. (2018) Values and Data Collection in Social Research, *Philosophy of Science*, 85, (1), 144-163.