Can we rely on Randomized Field Experiments (RFE's) in policymaking?

Mechanical objectivity and non-epistemic values in the Finnish Basic Income Experiment

Matias Valtteri Frantsi

The University of Helsinki

Faculty of Social Sciences

Social and Moral Philosophy

Master's thesis

April 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Social Sciences	Koulutusohjelma – Utbildingsprogram – Degree Programme Social and Moral Philosophy

Tekijä – Författare – Author
Matias Valtteri Frantsi

Työn nimi – Arbetets titel – Title
Can we rely on Randomized Field Experiments (RFE's) in policymaking? Mechanical objectivity and non-epistemic values in the Finnish Basic Income Experiment.

Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track
Social and Mora Philosophy

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	April 2020	96

Tiivistelmä – Referat – Abstract

The Nobel Prize in Economic Sciences in 2019 was awarded to Banerjee, Duflo, and Kremer for their fight against poverty, as well as for their methodological contributions to development economics. This thesis discusses their methodological approach, the use of randomized field experiments (RFE) in policymaking, which according to the advocates of the evidence-based policy (EBP), provide better and more objective evidence. This claim will be examined, and rejected, in the light of the methodological literature of field experiments in economics and a case study of the Finnish Basic Income Experiment (BIE Finland).

It will be argued that EBP's view on RFE's objectivity is rooted on the narrow view of *mechanical objectivity*, which overemphasizes methodological norms, such as randomization. This hinders various value choices regarding the research process and ignores the fact that the quality and nature of the evidence can change in the process. Co-creation of the scientific methods and interaction of the science and policy, thus, challenge EBP to reconsider their normative guidelines.

This thesis examines BIE Finland and demonstrates how ethical values can become as constitutive values of the research via decisions over (i) the experimental design and (ii) theoretical content, and via (iii) interpretation of the results. It will be argued that these three routes present epistemic risks, but also opportunities to increase the relevance and validity of the research. Ultimately these routes show how scientists are troubled by uncertainty and the risk of error, providing also an avenue for subjectivity. While these routes show complex trade-offs between epistemic and non-epistemic values, their implications for the objectivity of the research are also not clear. This is not only because, as will be illustrated with BIE Finland, RFE's are compatible with various epistemic aims and inferences that are not always clear, but also because the consequences of *inductive risk* for the normative guidelines and evidential standards is neither obvious. It will be argued that EBP should clarify the constituents of the trained judgment and the role of epistemic and non-epistemic values throughout the research process, because it ultimately shows how researchers are troubled with uncertainty and the risk of error. This requires them to abandon the value-free ideal and move beyond narrow mechanical objectivity in order to address the epistemic risks and potential disappointment associated with the evidence-based policymaking.

Avainsanat – Nyckelord – Keywords
Randomized field experiments, evidence-based policymaking, mechanical objectivity, value-free ideal, methodology of economics

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Caterina Marchionni

Säilytyspaikka – Förvaringställe – Where deposited
Helda

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Valtiotieteellinen tiedekunta	Koulutusohjelma – Utbildingsprogram – Degree Programme Käytännöllinen filosofia	
Tekijä – Författare – Author Matias Valtteri Frantsi		
Työn nimi – Arbetets titel – Title Voimmeko luottaa satunnaiskontrolloituihin kenttäkokeisiin poliittisessa päätöksenteossa? Mekaaninen objektiivisuus ja ei-tiedolliset arvot perustulokokeilussa.		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Käytännöllinen filosofia		
Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year Huhtikuu 2020	Sivumäärä – Sidoantal – Number of pages 96

Tiivistelmä – Referat – Abstract

Vuonna 2019 taloustieteen Nobel-palkinto annettiin Banerjeelle, Duflolle ja Kremerille heidän pyrkimyksistään vähentää köyhyyttä sekä heidän kontribuutioistaan kehitystaloustieteen metodologiaan. Tämä pro gradu -tutkielma käsittelee heidän mentelmällistä lähestymistään ja erityisesti satunnaiskontrolloitujen kenttäkokeiden hyödyntämistä poliittisessa päätöksenteossa. Näyttöpohjaisen päätöksenteon kannattajien mukaan satunnaiskontrolloidut kenttäkokeet tuottavat parempaa ja luotettavampaa näyttöä. Tässä tutkielmassa tämä väite hylätään kenttäkokeita koskevan taloustieteellisen metelmäkirjallisuuden sekä perustulokokeilua koskevan tapaustutkimuksen nojalla.

Tutkielma osoittaa, että näyttöpohjaisen päätöksenteon käsitys satunnaiskontrolloitujen kenttäkokeiden objektiivisuudesta perustuu kapeaan käsitykseen mekaanisesta objektiivisuudesta, joka korosta menetelmällisiä sääntöjä ja normeja. Tämä kuitenkin piilottaa useat todelliset arvovalinnat tutkimusasetelman suhteen ja sivuuttaa sen kuinka näytön luonne ja laatu voi muuttua tutkimusprosessin aikana. Tutkijoiden ja päätöksentekijöiden osallisuus tieteellisten menetelmien yhteismuotoiluun haastaa näyttöpohjaisen päätöksenteon uudelleenharkitsemaan normatiivisia ohjenuoriaan.

Tämä tutkielma tarkastelee perustulokokeilua ja osoittaa, miten eettiset arvot voivat muuttua tutkimuksen perustaviksi ja välttämättömiksi arvoiksi (i) koeasetelman menetelmällisten valintojen ja (ii) teoreettisten valintojen kautta sekä (iii) tulosten tarkastelun ja tulkinnan kautta. Näistä kolme reittiä nostattavat tiedollisia riskejä, mutta myös mahdollisuuksia kohentaa tutkimuksen relevanssia ja validiteettia. Lopulta ne myös osoittavat sen miten tutkija suhtautuu ja asennoituu epävarmuuteen ja erheen mahdollisuuteen, mikä tekee niistä myös subjektiivisuuden näyttämöitä. Samalla kun ne havainnollistavat komplekseja vaihtokauppoja tiedollisten ja ei-tiedollisten arvojen välillä, niiden seuraukset objektiivisuudelle ovat myös epäselviä. Tämä ei johdu ainoastaan siitä, että satunnaiskontrolloidut kenttäkokeet ovat yhteensopivia useiden tiedollisten tavoitteiden ja päättelyiden kanssa, jotka eivät ole usein kovinkaan selviä, kuten perustulokokeilun suhteen osoitetaan, vaan myös siitä että induktiivisen riskin seuraukset objektiivisuudelle ja näyttöä koskeviin standardeihin ovat myös epäselviä. Tutkielmassa argumentoidaan, että näyttöpohjaisen päätöksenteon tulisi selventää inhimillisen harkinnan ja siihen vaikuttavien tiedollisten ja ei-tiedollisten arvojen roolia, koska ne osoittavat sen kuinka tutkijat suhtautuvat epävarmuuteen ja erheen mahdollisuuteen. Tämä edellyttää näyttöpohjaisen päätöksenteon kannattajia hylkäämään käsityksen tieteen arvovapaudesta ja rikastamaan käsityksiään objektiivisuudesta, jotta he voivat käsitellä satunnaiskontrolloituihin kenttäkokeisiin ja niiden näyttöön liittyviä tiedollisia riskejä ja pettymystä.

Avainsanat – Nyckelord – Keywords
Satunnaiskontrolloidut kenttäkokeet, näyttöpohjainen päätöksenteko, evidenssi, objektiivisuus, arvovapaus, taloustieteen metodologia,

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Caterina Marchionni

Säilytyspaikka – Förvaringställe – Where deposited
Helda

Muita tietoja – Övriga uppgifter – Additional information

Contents

1. INTRODUCTION 5
1.1 The context of the thesis 5
1.2 Overemphasis on mechanical objectivity 6
1.3 Co-creation of the RFE's 7
1.4 The contribution 8
1.4 The structure 9
2. EVIDENCE-BASED POLICYMAKING (EBP) 10
2.1 RFE's as policy experiments 10
2.2 EBP: Threats to objectivity 14
2.2.1 Distrust towards government agencies 14
2.2.2 Distrust towards fellow economists 15
3. ARE RANDOMIZED FIELD TRIALS REALLY OBJECTIVE? 19
3.1 Convergence of interests 20
3.2 Faithfulness-to-facts 24
3.3 Mechanical objectivity 27
4. THE BASIC INCOME EXPERIMENT AS AN ECONOMICS EXPERIMENT 33
4.1. Kinds of field experiments in economics 33
4.2. The Finnish Basic Income Experiment 39
5. TRAINED JUDGMENT AND VALUES IN RFE's 51
5.1 Beyond mechanical objectivity 51
5.2 The value-free ideal 54
5.2.1 Impartiality 55
5.2.2 Neutrality 59
5.2.3 EBP and values: ambiguities 62
5.3 Co-creation and the influence of values 65
5.3.1 Experimental design 70
5.3.2 Theoretical content 78
5.3.3 Interpretation of the results 82
6. CONCLUSIONS 87
References 89

1. INTRODUCTION

1.1 The context of the thesis

A significant effort of transforming development economics into experimental field brought the Nobel Prize in Economic Sciences in 2019 to Banerjee, Duflo, and Kremer (The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel 2019). This indicates a yet growing attention on randomized field experiments (RFE henceforth), which according to the academy, "now entirely dominate development economics" (Foreign Policy 22.10.2019; The Guardian 14.10.2019). While celebrating the Nobel Laureates mission of alleviating global poverty based on scientific evidence, the academy simultaneously legitimizes the current experimental approach to the Evidence-Based Policy (EBP). EBP is a view which encourages the use of rigorous evidence in policymaking and especially the use of RFEs over other available methods (Cairney 2016, La Caze and Colyvan 2017).

RFE's, also called randomized controlled trials (RCT), randomized experiments, or randomized evaluations, are field experiments that randomly assign the subjects to treatment and control groups and the treatment is believed to be effective if there is a difference in outcome between these groups (Duflo & Kremer 2005, pp. 205-206; Guala 2005; 63, 78-79; Reiss 2013, p. 174). RFE's allegedly provide better evidence that is more transparent, reliable, and unbiased causal estimate of the true impact of the policy program (Duflo & Kremer 2005, pp. 205-210; Faverau & Nagatsu 2020a, p. 10; Orr, 1999, pp. 2-10). The underlying idea of the randomization is that if perfectly conducted, it allows two situations to be exactly comparable despite variation of the known and unknown background factors, and thus the difference in the size of the effect can be imputed to the treatment alone (Guala 2005, pp. 63, 78-79; Rawlings 2005, p. 193).

I will follow Faverau & Nagatsu's (2020a) terminology of RFE that encompasses both social experiments and modern RCT's in development economics. Therefore, what have become labelled as the First and the Second waves of RFE's will be elaborated in this thesis (see de Souza & Leao 2019; Faverau & Nagatsu 2020a, pp. 8-11; Heckman 2019). My thesis can also help to clarify similarities and differences between the two waves of RFE's regarding their features and types of inferences as I will compare BIE Finland to both (see section 4.1).

I will discuss whether RFE's are a preferable method to guide policymaking due to their objectivity. I am relying on the methodological literature and the Finnish Basic Income Experiment as a case study. I will argue that the EBP has not established a sufficient ground for preferring RFE's due to their objectivity, because of two interrelated reasons. Firstly, EBP relies heavily on a narrow view of objectivity that emphasizes strict and insufficient methodological norms and protocols. I refer to this as overemphasis on mechanical objectivity. Secondly, because science-policy is increasingly being co-created by the researchers and policymakers in settings that cross institutional boundaries and traditional responsibilities, EBP needs to accommodate this new reality into the methodological guidelines and norms. I will refer to this as co-creation of the RFE's.

1.2 Overemphasis on mechanical objectivity

EBP seems to overemphasize a particular type of objectivity: *mechanical objectivity* (see also de Souza Leao & Eyal 2019, p. 390). Mechanical objectivity refers to the strict rule-following and methodological protocols underpinning research (Daston & Galison 2007, pp. 121, 256). These norms and protocols are especially associated with transparent deductive inference guidelines of RFE's (Cartwright 2007; Reiss 2014, pp. 137-138). Even though, other dimensions of objectivity have been recognized in the literature and connected with the RFE's (Reiss 2014; pp. 137-138), I think their importance and implications are much less discussed in the literature, especially in EBP. The notions of *considered judgment* (Reiss 2014) and *trained judgment* (Daston & Galison 2007, pp. 18-19) will be discussed later in section 5. Their importance has been recognized in the final stages of the research, in which data is being analysed and conclusions be drawn. However, as I will argue, this subjective dimension of human judgment goes also deeper into the experimental design and theory choice than randomistas are perhaps willing to accept. I will argue that these dimensions of human judgment raise the risk of subjectivity and undermine to some degree the claims concerning RFE's superior objectivity.

My approach is inspired by the risk account of objectivity (Hacking 2015; Koskinen 2018), which puts epistemic risks and vices of the research first. This account takes objectivity as a negative account of describing the absence of epistemically harmful features (Koskinen 2018, p. 17). Objectivity is not a positive quality, or specific virtues, but implies the absence of epistemic vices that we should focus on (Hacking 2015; pp.

22, 26). According to Koskinen, all applicable concepts of objectivity have identified one or more potential epistemic risks that derive from our nature as imperfect epistemic agents (2018, p. 17). Our reliance on X is due to our belief that these risks are "effectively averted" (ibid., p. 9). I will illustrate with BIE Finland how methodological guidelines and norms associated with EBP are insufficient, because they cannot ensure that our beliefs about epistemic risks are "effectively averted" (see Koskinen 2018).

1.3 Co-creation of the RFE's

I will argue that EBP has not sufficiently recognized various epistemic risks and opportunities deriving from the fact that RFE's and evidence-informed policy are being increasingly co-created by various actors (see section 5). I will argue that the value-free ideal, impartiality, and neutrality that EBP largely assumes, and which seem to coincide with randomista's implicit understanding of mechanical objectivity, are insufficient basis for objectivity in policymaking. Inspired by the philosophers of science, who have considered the role of values in the scientific process (Douglas 2000, 2009; Elliott and Richards 2016; Longino 1990; Risjord 2014), I will demonstrate how political and ethical values are manifested in BIE Finland at various stages of the research including experimental design, theory choice, and interpretation of the results. These present three distinct routes through which political and ethical values can influence and alter the research. I will conclude that they all present epistemic risks or opportunities for the research, making them relevant considerations regarding the evidential standards of the EBP. Therefore, they need to be more elaborated by EBP.

This confirms the arguments of Douglas (2000) that researchers are forced to consult ethical values at various stages of the research, because of possible non-epistemic consequences of the research. However, where Douglas (2009, pp. 44-86) appeals to moral responsibilities of scientists, I will argue that their role as a researcher also require them to consider ethical values beyond impartiality and neutrality, because failure to do so can have consequences for objectivity and authority of the research. However, more empirical evidence is needed regarding the potential value trade-offs and their implications for objectivity in various instances. I suggest that future research should focus on the expanding roles and responsibilities of scientists, on adaptability and

resilience of the scientific process, and on more elaborate accounts of potential epistemic and non-epistemic trade-offs.

1.4 The contribution

Some experimental economists have already recognized the need for economists' new skillsets and even for wider responsibilities when conducting field experiments while paying a much closer attention to pragmatic success factors (see Duflo 2017; List 2011). Also, philosophers of social sciences have started to rethink moral responsibilities of scientists and implications for the normative philosophy of science that derive from more realistic understanding of the science-policy interaction (see Douglas 2009, Ruphy 2006). This has required a better understanding of how values manifest in science and how researchers are deeply concerned with uncertainties when making decisions (see Douglas 2000; 2009; 2016; Elliott and Richards 2016; Stegenga 2016).

This thesis expands on these observations and challenges to ponder the role of values more explicitly in the context of RFE's. While I follow Douglas (2000; 2009) and Longino (1990) in arguing against strict distinction between epistemic and non-epistemic values, especially regarding their implications for objectivity. I maintain that they can provide a conceptual framework to clarify the potential trade-offs and their legitimacy in policymaking, on which we can build. As several philosophers have suggested moving towards coupled epistemic-ethical frameworks (Khosrowi 2018; Khosrowi & Reiss, 2019; Risjord 2014) and urged for "more sophisticated picture of 'properly conducted' science than the 'traditional one'" (Ruphy 2006, p. 212), I will suggest exactly the same. I will also provide such a picture in the context of RFE's and BIE Finland.

As a result of my conceptual work and the case study, I will contribute to what can be called as natural history of science (see Currie 2015, pp. 553-572) by documenting methodology of RFE's and BIE Finland from the conceptual point of view. I will also demonstrate the potential that RFE's can have in the policymaking and present a counterargument to the philosophical claim advocating the use of RFE's.

Regarding EBP, I will conclude that they need far more precise account of the epistemic and non-epistemic values that can potentially alter and influence, either legitimately or illegitimately, the scientific process. I suggest that this could be achieved by addressing

more explicitly the constituents of trained judgment. This requires one to go beyond mechanical objectivity and to reject the value free-ideal. Such approach could build on the expanding literature and case studies reflecting the role of values in science (see Elliott and Richards 2016). BIE Finland exemplifies the complicated dynamics of epistemic and non-epistemic values, causing threats and opportunities for the research. Even when these values play legitimate role in the research process, greater transparency is needed. This is because these value choices challenge us to ponder our expectations about science and the responsibilities of scientists, which, I assume, require more elaborate ethical and political discussion also.

While I do not think that the use of RFE's in policymaking is unsound or unjustified, I am genuinely sceptical about the simplistic models of policy-informing that presume distinction of labour between science and policy (see also Cairney 2016). I think there are more reasons to consider BIE Finland as sound science than unsound, but BIE Finland nevertheless illustrates dissatisfaction and disappointment that EBP must deal with.

1.4 The structure

This thesis is composed of six chapters. After the introductory chapter (Chapter 1), the next chapter (Chapter 2) introduces the idea of RFE's in policymaking and related controversies. Chapter 3 deals with objectivity of the RFE's and dismisses two naïve views of randomistas. It also illustrates how *mechanical objectivity* has become essential and overemphasized in EBP. Chapter 4 elaborates the different kinds of experiments in economics and compares BIE Finland to the first and second wave of RFE's, demonstrating various aims and epistemic risks underlying RFE's.

Chapter 5 dives into the core issue of this thesis and deals with the challenge of the co-creation of the experimental methods. In this section, I will elaborate the role of values in the scientific process and reject the value-free ideal as a correct representation of the values in science.

Section 6 concludes by summarising the main points.

2. EVIDENCE-BASED POLICYMAKING (EBP)

This chapter will introduce the idea of RFE's in policymaking and present a picture of how they stand in comparison to other available methods in economics. I will discuss two threats to objectivity that ultimately motivate EBP's preference of RFE's: distrust towards government agencies and distrust towards fellow economists.

2.1 RFE's as policy experiments

Economists have been using a great variety of methods for causal inference (List 2011, pp. 8-9). These methods include multiple regressions analysis, instrumental variables approach, natural experiments, a regression discontinuity approach, structural modelling, and propensity score matching (List 2011, pp. 8-9). Experiments do, however, have a special role among these methods as they are perceived as the best way to establish causal relations, providing an ideal for nonexperimental research also (Harrison & List 2004, p. 1009; Morton & Williams 2010, pp. 17-19).

Broadly speaking, one can distinguish between two alternative approaches to a causality based on the direction to which they proceed: estimating the effects of causes and understanding causes of effects (Morton & Williams 2010, pp. 33-35). Experimental research in economics is quite often referred as a former type of research, aiming at estimating causal effects (Gerber & Green 2002 in Morton & Williams 2010, p. 35; List 2011 p. 8).

Policy experiments are field experiments in which a government agency or other institution acts like an experimentalist and manipulates variables such as government policies (Morton & Williams 2010, p. 54). The aim of this type of research can be labelled as *policy evaluation* or *impact evaluation* (Ludwig et al. 2011, p. 3; Rawlings 2005, p. 193). The idea is the same: to experimentally evaluate the true impact of the policy with the help of convincing counterfactual (ibid.).

The recent literature has suggested dividing policy relevant RFE's into two waves (see de Souza Leao & Eyal 2019; Faverau & Nagatsu 2020a; Heckman 2019). Both waves deploy randomization and aim at informing policymaking (Faverau & Nagatsu 2020a), but the theoretical motivations, the scale of the experiment, the policy context, and underlying

social success-factors of popularity have varied (see Heckman 2019, pp. 7-10; de Souza Leao & Eyal 2019).

First wave of policy experiments, labelled as social experiments, were run to test social policies in the US starting in 1960's (Ferber & Hirsch 1978, pp. 1380-1381; Greenberg & Robins 1986, p. 341-342; Heckman 2019, pp. 4-5; Levitt & List 2009, pp. 2-5). More recently, in 21st century, the number of randomized experiments has increased dramatically as RFE's are extensively used to inform public policies in both governing developing countries and developed countries (Baldassarri & Abascal 2017, pp. 48-49; Heckman 2019; de Souza Leao & Eyal 2019, pp. 383-384).

J-PAL (2020) alone reports nearly 1000 trials, most of them conducted in developing countries. In developed countries, UK has been in the forefront of the experimental governance with the governmental units such as Behavioral Insights Team (BIT). France has witnessed similar trends although in a smaller scale (Faverau & Nagatsu 2020a). The government of Finland set experimental governance as one of their primary targets for reform in the government platform 2015-2019 and initiated a large-scale RFE to be run 2017-2018: the Finnish Basic Income Experiment (BIE Finland) (Hallituksen julkaisusarja 10/2015). More recently, it was announced that one of the strategic research themes of Academy of Finland for 2020 is on evidence-informed decision-making (Valtioneuvoston päätös VNK/2019/111), indicating a yet growing interest and funds.

The strongest advocates of EBP hold that RFE's produce better evidence than other available methods (Cairney 2016, pp. 1-3; de Souza Leao & Eyal 2019, p. 384; Duflo & Kremer 2005, pp. 205-210; La Caze & Colyvan 2017, pp. 4, 10). While recognizing that RFE's are not suitable to address all issues (Duflo & Kremer 2005, p. 205) and that complementing evidence and a series of experiments are often needed (Banerjee & Duflo, 2011, pp. 14-15), many have started to think field experiments, especially RFE's, as "gold standards" to find out *what works* (see Baldassarri & Abascal 2017, p. 49; Cartwright 2007; Orr 1999; p. 2; Ross 2013). Over the past few years the excitement has perhaps become more moderate and economists have adopted more nuanced views on evidence (Dreze 2018, p. 45). Yet, as the recent Economics Nobel prize indicates, optimism remains, and the excitement is not entirely gone.

The advocates of the EBP tend to think that RFE's are especially useful for guiding public policies due to their main characteristics such as randomization or the lack of theoretical

assumptions (Bossuroy & Delavallade 2016, p. 149). According to Duflo (as cited in The Guardian 14.10.2019), randomistas do not aim at studying the deep roots of the interconnected societal issues, but "unpack the problems one by one" to see "what works, what doesn't work, and why". The ethos of RFE's claiming to find out *what works* often relies, at least implicit and also disputed, analogy with clinical trials used in medicine, which have for a long time been used to test effectiveness of the drugs[1] (Banerjee & Duflo 2011, p. 8, 2019, p. 8; La Caze & Colyvan 2017).

According to another increasingly popular metaphor, economists are like plumbers, who use a combination of trial and error, scientific intuition, and professional expertise to find answers to the smaller and more manageable problems than grand economic theories (Banerjee & Duflo 2011; p. 8; 2019, p. 7; Duflo 2017). They are "more concerned about 'how' to do things than about 'what' to do" (Duflo 2017, p. 3). Yet, RFE's are the primary tool of economist-plumber (Duflo 2017, p. 16).

The EBP and the related hierarchies of evidence emphasizing RFE's, have also caused a serious pushback in the academia against "methodological intolerance" of the EBP and "hyper-empirism" (Harrison 2013; Ross 2013). Some authors have strongly argued against what they take as a mistaken identification of field experiments with random sampling and disconnect from theory (Harrison 2013; 2014).

The view of economist-plumbers as good and privileged advisors of public policies has also been questioned (de Souza Leao & Eyal 2019; Dreze 2018, pp. 46-47;). Dreze (ibid.) is sceptical about economists being well qualified to design and influence policy as they often lack relevant competences regarding the details, such as programme implementation.

Powerful criticism has also been expressed in the public before and after the nomination of Nobel Prize in Economics to Banerjee, Duflo, and Kremer in 2019, questioning the epistemic authority and prominence of RFE's on the grounds of lacking insights, credibility and adequacy (see Foreign Policy 22.10.2019). According to this criticism, RFE's only validate common sense, cannot reveal causal mechanisms, or tell how to design complex institutions (ibid.). Furthermore, they are neither helpful for extrapolation nor understanding different segments of population (ibid.).

[1] This analogy has been contested also (see Faverau 2016; La Caze & Colyvan 2017;).

In 2018, 15 economists including former Nobel laureates had sent a letter to British Guardian expressing their worry that relying on RFE's will "lead to short-term, superficial, and misplaced policies" on aid and welfare spending (Business Standard 15.10.2019). These economists were particularly worried about ignoring broader macroeconomic, institutional, and political causes while evaluating only smaller short-term micro-interventions, often with high economic expenses (ibid.). Similar observations about the nature of the evidence provided by the modern RFE's have also been expressed in the published journal (see de Souza Leao and Eyal 2019, pp. 397-398).

Duflo (2019) addressed this critique in her Nobel lecture with the example of micro loans. She argues that their critics misunderstand how policies and research interacts and how the former is being influenced by the latter. According to her narrative, some of the early experiments concerning micro finance were promising, which resulted in positive media coverage on the topic. But as the experiments were replicated elsewhere, researchers found more controversial evidence. For a long time, researchers were not able to give straightforward recommendations for either side, because neither side had enough evidence backing up their claims. According to Duflo, the effects were heterogeneous and mediocre. In addition, researchers were not working only on one type of microfinance product but on several different products, which made it harder to make convincing comparisons. Only as studies accumulated, meta-analyses were able to demonstrate inefficiency of micro loans.

Given the fact that RFE's can in some instances, even according to randomistas, produce misleading result, one can wonder why we should trust in their evidence and rely on it in the policymaking. Also, the fact that reliability and usefulness of the RFE's have been contested both in the academia and in the public, highlights the urgent need to comprehend better the epistemic authority of RFE's in policymaking.

In the following section, I will illustrate two broad threats to objectivity that motivate the EBP's push for a greater use of the experimental evidence in policymaking. First is distrust towards policymakers and the lack of evaluation by the government agencies. This motivates the EBP's prior claim (see Cairney 2016; La Caze & Colyvan 2017) that evidence should be used more in the policymaking.

Second threat of subjectivity is distrust towards fellow economists, and especially towards economic theories and predictions. It is also clearly manifested in the arguments

of randomistas. This view, ultimately, perceives science and scientific heterogeneity as a source of variation and a threat of subjectivity. It motivates the EBP's second claim (see Cairney 2016, La Caze & Colyvan 2017), which advocates for a certain type evidence: evidence produced by RFE's. This provides a background to better understand the context of the EBP.

After dismissing two naive and simplistic views about objectivity of the RFE's (3.1-3.2) and illustrating the type of objectivity that underpins the EBP (3.3), I will discuss the kinds of field experiments in economics and compare BIE Finland to two waves of the RFE's (chapter 4). I will argue that it is not entirely clear what counts as a successful RFE. In chapter 5, I will discuss a challenge to objectivity of the RFE. I will argue that co-creation of the experimental methods raise subjective threats and epistemic risks that EBP must recognize.

2.2 EBP: Threats to objectivity

In this section, I will demonstrate how the EBP is characterized by distrust towards government agencies. The main reason for this distrust is that government agencies do not evaluate their policies. I will also illustrate how the EBP is characterized by distrust towards fellow economists who rely on other, less reliable, methods or on economic theories. These other methods, according to randomistas, are more likely to suffer from subjective biases and thus they present threats to objectivity. I will not discuss justifications of these claims any further.

2.2.1 Distrust towards government agencies

Donald T. Campbell published his famous article *Reforms as Experiments* in 1969, where he argued for the systematic experimental approach to the program evaluation of the government policies. According to him, this should be done by experimentally assessing the outcomes[2] and effectiveness of policy alternatives. He notices that social reforms are often advocated as if their success were certain.

[2] By outcomes I will refer to the changes in the variables brought by the manipulations of the treatment variable. Effectiveness refers to the ratio between the treatment variable and the measured outcome variable. The impact refers to the difference between the treatment and the control group and is

Campbell (ibid.) suggests that instead of persistently advocating for a specific reform, the government agencies should instead address the problem seriously and explore the alternatives on an experimental basis. Campbell also notices that evaluation is often restricted to one policy, whereas in the experimental ideology the data would be gathered on the designated control groups as well. Campbell discusses different types of quasi-experiments as well as true experiments that deploy randomization and control groups and argues that almost always true experiments should be preferred over quasi-experiments if both are available (p. 426).

More recently, but in a similar spirit, Banerjee & Duflo (2011, p. 16; 2019, p. 326; Duflo 2017; p. 9) repeatedly argue that policies often fail for three reasons: ideology, ignorance, and inertia. They claim that policies are often promoted either by the experts, the aid workers, or the policymakers, for ideological reasons, by ignoring the reality of the field (Duflo 2017, p. 9). And once the policies have been decided and designed, there is little room to evaluate and change them (ibid.). Duflo and Kremer (2005, p. 206) also notice that: "All too often development policy is based on fads, and randomized evaluations could allow it to be based on evidence".

Distrust towards government agencies and policymakers has thus been an important motivation for the EBP. According to de Souza Leao & Eyal (2019, p. 389), the EBP's new push for the RFE's attacked the older foundations of the development aid that relied on the managerial expert judgment. This expert judgment had to a large degree coincided with the liberal free market doctrine of the Washington consensus (ibid.).

2.2.2 Distrust towards fellow economists

De Souza Leao and Eyal (ibid.) also notice that discussions about causal identification within the economics profession gave an opportunity for the younger economists to shield themselves with the help of RFE's against the internal struggles of the discipline. These discussions about causal identification "disrupted business as usual and destabilized disciplinary objectivity" (ibid., p. 406). This relates to the distrust towards applied econometrics in 1980's, which was an important feature of the credibility revolution in economics (Angrist & Pischke 2010; de Souza Leao & Eyal 2019). The credibility

calculated from the difference in effectiveness in these groups. I believe this is the conventional usage of the terms (see Orr 1999, p. 3).

revolution paved way for a better causal identification strategies, especially design-based studies and experimental methods (ibid.).

According to the standard story of credibility revolution, much of the applied econometrics lacked credibility in 1980's due to the insufficient research designs and a lack of foundation for causal conclusions (Angrist & Pischke 2010, pp. 8-9). Even though many great studies were conducted at that time, much of the applied econometrics were conducted without providing a substantial justification for the assumptions (ibid., 8, 11). The use of instrumental variables, for instance, "was typically mechanical, with little discussion of why instruments affected the endogenous variables of interest or why they constitute a 'good experiment'" (Angrist & Pischke, 2010, p. 8). As a result, trust to the research decreased, and Leamer famously described the predicament: "hardly anyone takes anyone else's data analysis seriously (Leamer 1983, p. 37 as cited in Angrist & Pischke 2010, p. 3).

Duflo and Kremer also discuss the evaluation problem in the field of development (Duflo & Kremer 2005, pp. 206-208). The evaluation problem arises from the counterfactual question. What are the effects in the absence of the program and what is the potential influence on the subjects not participating to it? Duflo & Kremer argue that comparison over time does not provide a reliable estimate because other variables may not have been constant. Simple comparison between those who participated and those who did not, works neither, because it is a subject for selection bias, and hence does not account for potentially existing difference between two groups. The solution is to randomly select the subjects to treatment and control groups, whether constituted in individuals, communities, schools, or classrooms.

Other advocates of the social experiments have also emphasized the limitations of these other methods, including pre-post designs and comparison groups (see Orr 1999, pp. 4-9). They have argued that random assignment is a better way of constructing counterfactual in the program evaluation (ibid.).

Duflo & Kremer also discuss other methods that can be used to solve the selection bias and the problems with omitted covariates (2005, pp. 208-210). They discuss propensity score matching and argue that it is overdemanding regarding its demand to identify all potentially relevant differences between two groups. Difference-in-difference brings into the analysis untestable identification assumptions and can be biased by time-persistent

shocks or the existence of other implemented programs. Regression discontinuity design also faces problems with the identification. Overall, these methods require a greater attention on the identification strategies and "are less transparent and more subject to divergence of opinion" than randomized evaluations are (p. 210).

RFE's can therefore be contrasted with the non-transparent methods in applied econometrics that were lacking explicit causal identification, as well as simpler methods used for programme evaluations such as time and group comparisons. Distrust towards economics is also manifested in other popular argumentative strategies.

The advocates of the RFE's often highlight fallibility and biasedness of economic theories, predictions, and expert judgments (de Souza Leao & Eyal 2019, p. 409). Banerjee & Duflo (2019, pp. 6-8) argue that economists and economic theories often get things wrong. They attack bad economics and predictions, as illustrated by IMF's forecasts, which for two year-periods between 2000 and 2014 were "about as bad as assuming a constant growth rate of 4 percent" (The Economist as cited in Banerjee & Duflo, 2019, p. 6).

Banerjee & Duflo (2019, pp. 2-7) argue that in the eyes of the public, economics is also losing its authority. They report a significant distrust towards economists via a series of surveys in UK and US, illustrating that very few laymen trust economists even about their own field of expertise. According to them, only 25 percent of people trusted economists, which is only slightly better than politicians. People were also rarely willing to change their minds based on the reported consensus of the economists when disagreeing.

This sceptical stance towards theorizing and predicting is perhaps in line with the consensus of the economics discipline. Given the discipline's failure to predict 2008 financial crisis, many economists seem to become more aware of the limitations of economic theories and predictions. Nevertheless, the armchair economists are contrasted with the virtuous experts represented by EBP, who do not have many opinions to start with, as they climb down from the ivory tower to do serious legwork (de Souza Leao & Eyal 2019, p. 409).

Development economics has also witnessed a lively debate whether researchers can do without almost no theory (Ross 2013, p. 127). The idea associated with RFE's is that experimental data should speak for themselves without unnecessary strong assumptions (Banerjee & Duflo 2011; Gerber & Green 2012 as cited in Ross 2013). This position can

also be labelled as *Hyper-Empirism*, which takes randomization as the gold standard for identification of the causes and magnitude of the causal effects (Ross 2013, p. 128).

According to Duflo (2017, pp. 3-4), theoretical assumptions rarely hold in the real world and models yield significant uncertainties. Intuition even when grounded in theory and existing evidence is not sufficient for policymaking, but "a very poor guide of what will happen in reality" (Duflo 2017, pp. 4, 15). RFE's are better, because they simply have less assumptions (Bossuroy & Delavallade 2016, p. 149).

Many have thus acknowledged that the grand theoretical frameworks in economics have inadequately implied assumptions about homogeneity of responses to incentives, while heterogeneity might better describe the applied situations (Ross 2013, pp. 127-128). This argument casts doubt on the grand theories and explains partially why randomized field evaluations have gained popularity. Another reason is the impressive growth of the computing power that have led to the sophistication of the econometric tools, revealing even more possible sources of estimation bias.

EBP and RFE's allegedly provide an answer to the lack of evaluation by the government agencies and to the bad economics, both threatening objective policy evaluations. Their approach relies on RFE's, which allegedly are less prone to subjective biases, and can help policymakers and researchers to put facts and evidence first.

3. ARE RANDOMIZED FIELD TRIALS REALLY OBJECTIVE?

In this chapter I will discuss objectivity of RFE's. I dismiss two naïve arguments about objectivity of the RFE's that can be associated with the EBP. The first is that RFE's are great dispute-reconciliation mechanisms as they demonstrate convergence of interests. The second is that the experimental method reveals brute facts about the world as it is, which is the very reason that RFE's provide objective evidence. Finally, section 3.3 demonstrates how *mechanical objectivity* has become essential and, perhaps, overemphasized in the experimental economics.

We have seen that part of the push towards the use of RFE's is meant to address the threats of subjectivity in policymaking. Yet *objectivity* has various historical and conceptual roots and it is a vague and loose concept (Daston & Galison 1992, p. 82; Douglas 2004, p. 453, 2009, pp. 115, 129; Hacking 2015, p. 19). Objectivity, especially in the modern usage, implies both intersubjective and evaluative meaning (Hacking 2015, p. 22). It has a rhetorical implication that "I endorse this, and you should too" (Douglas 2004, p. 453).

Objectivity implies a contrast with one's subjective personal viewpoint (Daston & Galison 1992; pp. 82, 84; 2007, pp. 36-37; Hacking 2015; p. 21-22). This subjective viewpoint can be mistaken, or partial in a sense that it does not characterize the whole truth (ibid.).

In the following two sub-sections, I will discuss two assertions relating with objectivity of the RFE's that can be associated with the EBP. The first assertion is that researchers and policymakers should put facts and experimental evidence first, because this type of evidence is objective in a sense that it represents the world more correctly. The second assertion is that RFE's are good, because they provide good dispute-reconciliation mechanism for policymakers and researchers. I will argue against both assertions. In the third sub-section, I will discuss a more adequate way to think about objectivity of the RFE's, mechanical objectivity, which I take it to be the main, but not only, account of the objectivity underpinning the EBP.

3.1 Convergence of interests

Not only the proponents of EBP suggest that experimental evidence can alleviate disputes over competing theories and assumptions, but they have also argued that the experimental method itself can bring scientists and policymakers together despite of their diverging interests and motivations (see Bossuroy & Delavallade 2016, pp. 147, 150-151). According to them, RFE's can foster convergence between policy design and economic theory (ibid.). This convergence of interests is manifested in studying causal mechanisms and it can allegedly provide valuable insight for both: theory-building and designing policies (ibid.).

Bossuroy & Delavallade (2016, pp. 150-152) argue that convergence between policy design and economic theory occurs when theoretical assumptions and behavioural predictions are designed within theories of change and logical frameworks. Logical frameworks are analytical policy tools of development aid, used to map out and communicate the logical connections of the programme, its activities, expected outputs and outcomes as well as causes and effects (The World Bank 2005, pp. 13-18). Theories of change are closely related as they fulfil the key assumptions and causal chains to the project design (White 2005, p.47).

Generally speaking, the idea of the logical frameworks is to ensure that the activities are aligned with the official thematic areas and aims, to ensure sufficient planning and monitoring, and to support coordination of the activities. One could consider it as a method connecting the UN sustainable development goals with the concrete activities of the NGO's, keeping the NGO's accountable.

Sometimes logical frameworks and theories of change are used mainly for mapping out the activities and the expected outcomes, but they can include further behavioural assumptions (Bossuroy & Delavallade 2016, pp. 150-152). Via RFE's policymakers can test their reasoning and the theories of change, where the inference is the weakest or the most uncertain, and researchers can test their theoretical models, which "crystallizes this convergence of interests" (p. 150). This approach is also sometimes called Theory-Based Evaluation (TBA) (White 2005, p. 47). Bossuroy & Delavallade (2016, p. 151) also argue that policy-informing is an iterative process of using RFE's and economic theory to open and close knowledge gaps.

This argument for the convergence of interests has not been well analysed and studied in the existing literature. Bossuroy & Delavallade do not give detailed support, empirical or conceptual, for this argument. Other writers have rejected the idea of RFE's as dispute-reconciliation mechanisms from the straight hand as completely uncritical hype of the RFE's but have neither discussed the issue in detail (Deaton & Cartwright 2017, p. 10).

I think the convergence of interest argument can be interpreted in two ways. Either the research process of conducting RFE is what brings policymakers and researchers together, uniting them by their common interest in the scientific *process* as it serves the purposes and interests of both. Or the experimental evidence, provided by the RFE's, is useful and valuable *output* for both and thus, it is in their common interest. While the former emphasizes the joint process as a uniting factor, the latter emphasizes the evidence and its usefulness as a common interest.

Despite recognizing the different perspectives and missions of the researcher and policymaker, Bossuroy & Delavallade seem to suggest both interpretations of the arguments. While researcher is interested in theory-building and policymaker in designing effective policy, they both are interested in testing specific theoretical assumption and the most uncertain parts of the logical framework (see Bossuroy & Delavallade 2016, pp. 150). As a result, researchers can build better and more accurate theories, and policymakers can design better policies based on the same piece of evidence.

I am highly sceptical about both versions of the argument. Let me start with the second interpretation.

Firstly, I think that Bossuroy & Delavallade ignore the fact that the evidence provided by the RFE's can be disputed and it does not necessarily serve the interests of policymakers and researchers equally. I think this is the case with the BIE Finland also. I will discuss the different aims that RFE's can serve in more detail later. It should nevertheless be noted here that despite experimental theory-building and policy design have something in common (see Roth 1986, p. 266), they might imply different experimental designs and features. Even though RFE's would be able to produce evidence of causal mechanisms, as Bossuroy & Delavallade put it (2016, p. 151), the usefulness of the evidence for policy might be to a much lesser degree and decrease in use (Deaton & Cartwright 2017). In the BIE Finland, the research group had also higher hopes for the strength of the evidence

produced, for instance, regarding the sample size or the replication of the experiment with varied level of basic income (Kangas 2016; Kela 2016, pp. 58-62).

Also, there is an argument that can support either interpretation. Some writers have suggested that RFE's can provide an ideal tool to convince other people with the diverging views and veto power (Banerjee et al. 2016 as cited in de Souza Leao & Eyal 2019, p. 408). Bossuroy & Delavallade (2016 p. 149) argue that simple evidence comparing control and treatment groups makes it "easily understandable and compelling to non-expert audiences". Deaton & Cartwright (2017) also notice that RFE's can be helpful when trying to persuade and convince non-expert audiences.

I do not consider this as convincing argument for either interpretation. As Reiss & Sprenger (2017) notice, clever marketing strategies can also promote trust in certain propositions. It still does not make the proposition any more objective (ibid.). We simply demand more from objectivity than fulfilling the condition of promoting trust or convincing the others (ibid.).

Moving to the first interpretation, BIE Finland also fails to support the argument that conducting RFE's will unite the policymakers and researchers by their common interest in the scientific process or in the specific issue. It is obviously true that both were interested in studying the labour market effect of the basic income in the Finnish context to some degree. But one must only read the bill about the basic income experiment in Finland to realize that even the aims of the experiment were rather vague and debatable. The bill includes multiple aims for the experiment: to "reform social security, to encourage participation and employment, to reduce bureaucracy, and to simplify the complicated benefits system in a sustainable way regarding public finances" (Kela 2016, p. 58). It is questionable how all these aspects could have even been incorporated into one study setting. This is probably why in the postscript, published by the research group, it was added that primary aim of the experiment was to promote employment (Kela, p. 58). Rather than carefully thinking how to achieve these broad and, perhaps impossible, aims of reducing bureaucracy, increasing employment, and meeting the needs of the changing work-life, the process seems to be more like narrowing down to a single question that can be studied via RFE. Notice how some authors have argued that RFE's are suitable to address only very narrow questions (Cartwright 2007; Reiss 2014), which

can lead to somewhat problematic changes in the research questions (Reiss 2014) or decreasing usefulness of the evidence (Deaton & Cartwright 2017).

This casts a doubt on how precise theories of changes and logical frameworks were even considered in the BIE Finland. The research group was given a comprehensive task to study four different models of the basic income, the taxation models for them, propose methods to integrate various social security benefits into the basic income, and assess the models (Kangas, Simanainen, Honkanen 2017, p. 88). It is safe to say that the focus was much on the institutional setting and figuring out how to finance the proposed basic income, to which purpose microsimulations were also conducted (see Kela 2016, pp. 16-49). The pre-study also compared the models based on their pros and cons concerning the aims (economic incentives for work, bureaucracy and administration, and poverty, income gaps and participation) and feasibility (feasibility, feasibility for testing, and support base) (ibid., pp. 53-57). But even them, it seems that feasibility, administrative considerations, and costs, played a major role in determining which model was to be tested (ibid., pp. 60-62). Interestingly all four models were identified mainly for their disadvantages regarding feasibility and feasibility for testing (ibid., pp. 56-57).

So, the pre-study did not elaborate possible behavioural responses or theoretical assumptions that could have guided the approach and helped to achieve these broad aims more rigorously. There seems to be little consideration regarding other measures that economic incentives to achieve those aims. Neither were behavioural assumptions concerning risk attitudes or time preferences, for example, considered. Therefore, the comparison did not identify the weakest and most uncertain parts of the reasoning, at least regarding the aims and theoretical assumptions of the experiment, which policymakers and researchers would have had common interest to study. Ant the process did not illustrate the convergence of interests. Quite the opposite.

Notice also how the head of the research group, Olli Kangas, demonstrated frustration with the planning process in his public lecture "How to plan a successful field experiment – and how to destroy it"[3]. He was especially frustrated with the planning process, negotiation with the policymakers, the budgets, and the schedule. He also noticed that the

[3] 13.8.2018, Public lecture, The University of Helsinki, Swedish School of Social Sciences, Festhall.

ministers of the three different political parties involved in the planning all had quite different views of the basic income and the respective experiment.

I think this shows rather convincingly that the argument for the convergence of interests requires quite more support than what was here discussed. It seems that designing and conducting RFE's, in fact, does not always crystallize the convergence of interests[4], as Bossuroy & Delavallade put it (2016, p. 150).

3.2 Faithfulness-to-facts

Among the advocates of EBP (see Banerjee & Duflo 2011, 2019; Campbell 1969) it is popular to argue that we need more facts and evidence on policies and *what works*. In empirical sciences also, researchers can be tempted to solve the disputes over competing theories by mere observation and experimentation, by discovering the brute facts (Reiss & Sprenger 2017). A view holding that objectivity derives from the correct representation of the world as it is, is called as *faithfulness-to-facts* (Reiss & Sprenger 2017). In this section, I will follow Reiss and Sprenger (ibid.) and reject faithfulness-to-facts as a correct view on objectivity.

The main reason to reject faithfulness-to-facts is that it is not applicable concept of objectivity, because it hides the choices and variability of the scientific process. It is not evidence or brute facts, per se, that are objective and give science its authority, but the type of process that underpins the production of it (Reiss & Sprenger 2017). Even though experiments are great method to gather empirical evidence, it is much harder to argue that this type of evidence is objective and aperspectival (Reiss & Sprenger 2017).

Reiss & Sprenger (2017) make a distinction between *process objectivity* and *product objectivity*. Product objectivity refers to the objectivity of the outputs that scientific endeavour produces such as scientific results, claims, theories and evidence[5]. Process

[4] De Souza Leao & Eyal (2019) have argued that popularity of the RFE's in development economics is, in fact, not explained by the better experimental designs, but by enabling factors that allow RFE's to function as "hinge" between two different fields and overcoming resistance. Their sociological analysis points towards variables, such as networks of expertise, short-term nature and fragmentation of the development aid, and diverse funding sources.

[5] It makes a difference how one characterizes these products and whether one talks about results, claims, theories, or evidence. Despite of being closely connected terms, they do not necessarily refer to the one and same thing.

objectivity refers to scientific process that produces those outcomes. Reiss and Sprenger argue that objectivity must be a feature of the scientific process, not its products.

So, let us look at the problems with faithfulness-to-facts view in detail.

First, observations are theory-laden in a sense that they are dependent on the terminology deployed by the researcher. (Kuhn 1962 [1970], as cited in Reiss & Sprenger, 2017). Researchers will select the key concepts according to the paradigm they are working with, and the meaning of the concepts can vary from paradigm to another. Like the meaning of the deployed concepts, perception of the researcher is also dependent on the paradigm that researcher is working with. Researchers will not necessarily see the same thing, even though they looked at the same thing from the same perspective.

In addition, scientific theories are rarely if ever measured against the pure observation (Reiss & Sprenger, 2017). Rather, they are tested against experimental facts, which are products of scientific measurement and experimentation. Researchers must make various choices regarding which instruments and measures to use. Different instruments and measurements might yield different results and interpretations.

Let me demonstrate these points with The Finnish Basic Income Experiment. How can we say that the evidence was faithful to facts or not? This is a question about whether or not the basic income unambiguously yielded employment effect and increased the labour market supply, by simultaneously providing better financial incentives and reducing bureaucracy gaps. So, why is this view on objectivity problematic?

Firstly, as all experiments, it included various conceptual choices such as *universal* and *unconditional basic income, labour market supply,* and *subjective well-being*. For example, the fact that the *child increases of the unemployment benefit* were not included in the basic income made the research group doubt that the tested basic income was not truly *unconditional* for families with children (Kangas et al. 2019, p. 13). This implies that conclusions about unconditionality could not be drawn in that regard.

As many commentators noticed on social media, the preliminary results of the Basic Income Experiment were used on social media by the public to support many different conclusions[6]. For instance, the government's institute for economic research (VATT)

[6] For instance, in social media, some commentators concluded that the experimental fact concerning the absence of the remarkable difference between the employment levels of the treatment and control

concluded that economic incentives did not make the unemployment benefit recipients more passive, because the experimental subject population did not avoid activation measures during the trial, such as personal meetings with employment officials (Hämäläinen et al. 2019; Yle News 2.4.2019). However, economic incentives did not make targets more active either in terms of having a significant employment effect (ibid.). Again, these are two observations that can be supported with the same experimental findings.

The research group also notices that even though the register data of the Finnish Tax Administration and The Finnish Centre for Pensions (Eläketurvakeskus) are available and can be utilized, "measuring employment on the basis of register data is not straightforward" (Kangas et al. 2019, p. 11). One of the challenges was to estimate the periods of employment and unemployment for the people having signed so-called zero-hour contracts (ibid.). Although the research group tackled this issue by using several measurements[7] (ibid.), this demonstrates the choices that researcher must make in the process regarding the instruments and measurements.

Estimating the periods of employment and unemployment with different instruments, measurements, and data sources is not without importance. For instance, Barnow & Greenberg (2015) have analysed a small sample of randomized controlled trials using earnings-related impact estimates and argued that different data sources can yield very different results. According to them, there is a significant difference in the size of impact estimates depending on the type of data source. They suggest that there is a trade-off between administrative data and survey data reporting the subjects' earnings as the administrative data is cheaper but might not include all income whereas the survey data usually report higher earnings.

It follows that experimental evidence, per se, cannot be considered as objective in the sense that it discovers brute facts and is faithful to facts. Conceptual choices, the fact that perception of the researcher is tied to the paradigm they are working with, and various

groups during the first year, showed ineffectiveness of the monetary incentives to affect unemployed people altogether. Alternatively, the same experimental finding was possible to interpret by the public as demonstrating the ineffectiveness of the proposed Basic Income (treatment group) or demonstrating the ineffectiveness of the current social welfare system relative to the proposed Basic Income model (control group).

[7] More specifically: "employment spells in the open labour market which exceed a certain wage level, the share of persons who have had any earnings or income from self-employment during the year, as well as total earnings and income from self-employment" (Kangas et al. 2019, p.11).

instruments and measures scientists use, all indicate that research is not objective due to faithfulness to facts but rather perspectival. However, this is not to say that the research could not be objective, only that objectivity must have another source than faithfulness to facts to have any applicability. The same applies to the Basic Income Experiment: conceptual and measurement choices and interpretation of the results, do not undermine the objectivity of the experiment, but suggest that we should move beyond product objectivity.

3.3 Mechanical objectivity

In previous section, I have dismissed faithfulness-to-facts, based on its commitment to product objectivity. In this section, I will discuss mechanical objectivity, which is a specific type of process objectivity. While process objectivity can refer to various aspects of the scientific process, such as measurement procedures, individual reasoning processes, and the social and institutional dimensions of science (Reiss & Sprenger 2017), mechanical objectivity also encompasses various dimensions of objectivity (Douglas 2004, p. 466). According to Douglas (ibid.), the original usage of the term (see Daston & Galison 1992), covered multiple dimensions such as procedural rule-following and personal restraint of using values in place of evidence[8].

This section illustrates how mechanical objectivity appears in the context of EBP as a specific type of methodological rule-following tied to the methodology of RFE's. I will illustrate how mechanical objectivity has become overemphasized by the EBP through the credibility revolution and the experimental turn.

Daston & Galison (2007, pp. 121-123, 256) describe mechanical objectivity as a strict rule-following and reliance on protocols that would produce results almost automatically.[9] These protocols, including statistical methods and experimental

[8] Douglas (2004, p. 466) labels three dimensions in this respect: procedural objectivity, detached objectivity, and convergent objectivity. She argues against expanding detached objectivity to value-free and value-neutral objectivity and notices that procedural objectivity does not guarantee either value-free or value-neutral objectivity (p. 466).

[9] Daston & Galison distinguish three eras of objectivity (Galison 2015, p. 58). Starting from the eighteenth century, the first era emphasized idealized universal objects and truth-to-nature. Depicted objects were not particulars but universals in their perfection and scientist had "the capacity to see behind the curtain of appearances" (p. 58). Truth-to-nature involved a metaphysical dimension and a desire to reveal a hidden reality, so difficult to obtain (Daston & Galison 2007, p. 58). It required sharp and sustained observation, patience, talent, and an ability to analyse and synthesize observations (ibid.).

protocols, aimed at reducing all the distortions deriving from the scientist's personal tastes, commitments, and ambitions. The protocols and machines that enabled scientists to see clearly did not express the freedom of will, but exercised freedom from will. "Machines were ignorant of theory and incapable of speculation" (p. 123).

Mechanical objectivity implied that idealization and aesthetic perfection became vices, as they were too subjective, and attending to particulars became a virtue (Galison 2015, p. 58). According to Daston & Galison (2007, p. 122) mechanical objectivity was strikingly distinct from any earlier efforts to capture nature right in its mechanical methods, restrained ethics, and individualized metaphysics. First being as a way of policing the artists, the mechanical protocols later became a mean of self-surveillance, as a form of ethical and scientific self-control (ibid., 174). Even though machines provided an ideal, it was clear that they did not necessarily guarantee a success as following of the protocols was sometimes extraordinarily difficult (Daston & Galison 2007, pp. 174-175).

Julian Reiss has explicitly connected the idea of mechanical objectivity with experimental methodology and the use of RFE's in economics (Reiss 2014, pp. 137-138). Reiss argues that the sense of objectivity that matters most in randomized controlled trials, and other methods that provide deductive proofs[10], is that the conclusions are established *mechanically* from the principles and assumptions of the method. The evidential relation between conclusion and assumptions is deductive and transparent. This mechanical process frees science from human judgments and from the personal elements that can bias the research.

> "What matters here is not so much that the conclusions are certain given the assumptions but rather they are established 'mechanically', with as little subjective judgments as possible. In consequence, results are established independently from the team of economists doing the experiment or derivation and in a way that is transparent to everyone who bothers checking them" (Reiss 2018, p. 138, original emphasis).

The second era, *mechanical objectivity,* started in the mid-nineteenth and it tried to eliminate desires and aims of the scientists, which were perceived inherently as subjective threats to objectivity (Daston & Galison 1992, p. 98; Galison 2015, p. 58). It was "self-denial coupled with the drive toward disciplined automaticity" (Daston & Galison 2007, p. 179).

[10] Reiss includes a broad range of methods from Mill's method of inference to the ideal RCT's, instrumental variables and other methods defined in "Mostly Harmless Econometrics" by Angrist and Pischke (2008) and methods that Cartwright calls as "clinchers" (see Cartwright 2007).

In fact, Reiss also discusses the problems and limitations of mechanical objectivity (Reiss 2014, pp. 140-141). He argues that mechanical methods are only suitable to produce answers to the narrow sets of questions. According to him, there are often troubling changes in research question from one that researcher wants to answer to one researcher can answer with the specified method.

Several other authors have also observed the connection between mechanical objectivity and EBP (see de Souza Leao & Eyal 2019, pp. 386-387, 409). There are also two clear historical reasons for the connection between RFE's and mechanical objectivity: the credibility revolution and the experimental turn in economics[11] (de Souza Leao & Eyal 2019, pp. 406-407). I will now discuss each one in turn.

The lack of trust in applied econometrics in 1980's was an important reason for better causal identification strategies to emerge through the credibility revolution (Angrist & Pischke 2010; de Souza Leao & Eyal 2019). Since then, the credibility of economics has been increasing for many reasons (Angrist & Pischke 2010, pp. 10-20). However, one of the main improvements of the credibility revolution was a greater emphasis on the *design-based studies*, of which experimentation and quasi-experimentation are part of (ibid., pp. 5-6, 26). Design-based studies have better prima facie credibility and a greater emphasis on the institutional and data-driven interpretation of causality (ibid., p. 5).

So, the credibility revolution was essentially an improvement in drawing causal conclusions, inspired by the ideal of a randomized experiment (Angrist & Pischke 2010, p. 12; Brodeur, Cook, Heyes 2018). Improved causal conclusions can be achieved with a variety of techniques including difference-in-difference (DID), instrumental variables (IV), randomized controlled trials (RCT), and regression discontinuity design (RDD) (Angrist & Pischke 2010, p. 12; Brodeur, Cook, Heyes 2018, p. 2). These methods seem to be united by their greater focus on research design and control on interfering variables, which is reflected on the ideal of as-good-as-randomly-assigned (see Angrist & Pischke 2010, pp. 12-13). Of these methods RCT's are sometimes thought to be the most trustworthy, which is manifested especially in the Evidence-Based Policy literature (for empirical argument favouring RCT's and RDD's see Brodeur, Cook, Heyes 2018).

[11] De Souza Leao & Eyal (2019) use the label of behavioural economics, but I think it is more accurate to talk of the experimental turn because it is not so much the content of behavioural economics but its method that is more clearly connected to mechanical objectivity.

Glenn Harrison (2013), however, has argued that there is a mistaken identification of the credibility revolution with randomization. Overall, there are two ways to establish causality in experimental reasoning: control and random assignment (Morton & Williams 2010, p. 26). Control and random assignment ultimately serve the same aim of isolating the causal mechanism and avoiding confounding factors, but in different ways (Morton & Williams 2010, pp. 141-142).

The idea of randomization is to randomly assign the units to the treatment and control groups, allocating evenly all the known and unknown confounding variables between two groups (Morton & Williams 2010, pp. 47-48). No matter if the confounding variables are observable or not, a sufficient sample size together with the central limit theorem ensures that two samples have the same distribution of properties (Bossuroy & Delavallade 2016, pp.148-149) Perfectly and ideally conducted randomization would lead to a situation, in which other variables influencing the measured effect will be evenly balanced between two groups, allowing researcher to measure the effect of the manipulated variable only (Morton & Williams 2010, pp. 47-48). In practice, randomization always involves various decision about what to randomize (ibid.). There also seems to be a great difference between ideal experiments and imperfect real experiments, as various untested assumptions are easily smuggled into the research design[12] (see Blalock 1991, pp. 331-332). Yet randomization is a powerful method to sidestep potentially troublesome variables without explicitly controlling them (Morton & Williams 2010, p. 101).

However, many authors have argued that the guiding research questions of the statistical treatment effect literature, which relies on RFE's, are very limited and narrow in scope (see Harrison 2013, Heckman 2019, Wilcox 2016). Wilcox (2016, p. 136-137) has criticized the EBP because there is often a greater interest in hypothesis-testing rather than measurement. He contrasts the lessons of the credibility revolution, namely randomization, with Nobel-winning experimentation in physics conducted by Millikan, showing how some of the most successful experimental work in natural sciences has nothing to do with randomization (ibid., for the same argument see also Harrison 2013, p.105). Wilcox (ibid.) also expresses his scepticism towards wide applicability and

[12] Blalock (1991, pp. 331-332) mentions manipulations and uncontrolled events during the experiment as well as matching techniques to compare target and control groups.

empirical success of hypothesis-testing, which ignores the art of measuring and conditioning the size of causal effects "on reliably measurable covariates" (p. 137).

Curiously, the credibility revolution was, on the one hand, about improving mechanical objectivity, despite the crisis of applied econometrics resulted at least partly from the mechanical rule-following, which lacked a rigour and causal identification. Nevertheless, the solution was to turn on better research designs, design-based studies, having more credible foundations for causality. As a result, new types of rule-following, associated with randomization, seem to emerge as being, perhaps, even more important than before.

Historically this period starting from the second half of 1980's also coincides with the experimental turn and the rise of experimental economics as a sub-discipline and methodological novelty in economics (see Svorencik 2015, pp. 10, 28, 238). The experimental turn also brings up constituents of mechanical objectivity. For instance, the experimental turn yielded a claim that rigorously collected laboratory or field data would be a better foundation for economic theorizing than other types of data (Svorencik 2015, pp. 14, 236). Svorencik distinguishes four essential drivers of the experimental turn: integrity, rigorousness, the virtuous cycle, and symmetry (Svorencik 2015, pp. 6, 238).

It seems that three out of four of these drivers can be understood as improvements in mechanical objectivity: advocating of the experimental data over other data sources (integrity), advocating the personal collection of data under controlled conditions (rigorousness), and advocating the equal appreciation of experimental data and theory (symmetry) (see Svorencik 2015, pp. 6, 238).

However, it is important to notice that mechanical objectivity is not the only way how objectivity can be understood. As Daston & Galison (2007) argue, different types of objectivity are not mutually exclusive, and they have been realized in varying degrees at different times in their study. This seems to be the case with credibility revolution and the experimental turn also. For instance, transparency, communicability, and more open discussion about causal identification strategies in economics, have seemingly played a major role in the credibility revolution (Angrist & Pischke 2010, pp. 12, 16-17). These collective aspects clearly go beyond mere individualistic rule-following, associated with mechanical objectivity. Also, as I will show later, *a considered judgment* (Reiss 2014, pp. 138-141) plays an important complementary role to mechanical objectivity that cannot be ignored.

Here, I have shown that the credibility revolution and the experimental turn in economics have rooted mechanical objectivity deeply into the core of the EBP. In next chapter (4), I will discuss the kinds of fields experiments in economics and contrast BIE Finland with the two waves of RFE's in economics. After that I will raise a challenge to the RFE's objectivity, which derives from co-creation (chapter 5). Chapter 5 opens a discussion about different values, whether scientific or political, and their implications for objectivity of the RFE's in policymaking.

4. THE BASIC INCOME EXPERIMENT AS AN ECONOMICS EXPERIMENT

In this chapter I will elaborate the kinds of field experiments in economics and compare BIE Finland to the first and second waves of RFE's in economics. The aim of the chapter is twofold: (i) to characterize aims and features of the BIE Finland and help reader to put the case study in the broader context of RFE's, and (ii) to illustrate ambiguities and value choices regarding what features and epistemic aims constitute a successful RFE. In section 4.1, I will elaborate the aims that RFE's can serve. In section 4.2, I will characterize BIE Finland.

4.1. Kinds of field experiments in economics

The methodological literature is only recently becoming aware of the complex methodological history of field experiments in economics and interested in clarifying it (see de Souza Leao & Eyal 2019; Faverau & Nagatsu 2020a; 2020b; Heckman 2019). Lately, other social scientists have also started to clarify the debates evolving around RFE's in their fields (see Baldassarri & Abascal 2017). However, much of the attention has been so far on a narrow issue of randomization as the advantages of field experiments go beyond mere randomization (Faverau & Nagatsu 2020a, 2020b; Levitt & List 2009, pp. 2-7).

In economics, the spectrum of the experimental methods and classifications regarding field experiments are broad, but borders between different types of experiments are often fluid and unclear (see Faverau & Nagatsu 2020a, 2020b; Harrison & List 2004). For instance, field experiments can be placed in between of laboratory experiments and observational studies, as they combine some aspects of both[13] (Boumans 2016). They can also be divided into sub-categories regarding the level of realism and natural environment they deploy (Faverau & Nagatsu 2020b, p. 11; Harrison & List 2004;). Types of field experiments can be classified by the type of control, and respective challenges regarding

[13] Realistic environment makes field experiments resemble observational studies and greater control makes them resemble laboratory experiments (Boumans 2016; Harrison and List 2004).

external validity[14] (Faverau & Nagatsu 2020a). Moreover, there are various types of policy experiments[15], not all being necessarily RFE's.

This experimental heterogeneity in economics as well as conceptual unclarities highlight an urgent need for case studies to better understand the potential that the RFE's can or cannot have in policymaking (see also Favera & Nagatsu 2020b). RFE's themselves can be used for very diverse inferences and to fulfil different epistemic aims, which have been too much overlooked in the literature[16] (Deaton & Cartwright 2017). According to the proponents of the RFE's, this flexibility can also be an advantage, which was reflected in Banerjee's (2019) Nobel lecture as he argued that RFE's are great for testing, unpacking, or building theories as well as testing and improving policies.

However, given the broad range of criticism (for summary of the literature see Faverau & Nagatsu 2020b, p. 2) we have reason to doubt that not all can be achieved at once. So, let me clarify different experimental aims that can be associated with the RFE's.

Alvin Roth (1986, 1995) have presented three, allegedly, useful metaphors to describe loosely different aims of experiments in economics: *speaking to theorists*, *searching for facts*, and *whispering in the ears of princes*. These metaphors were intended to help researchers to understand different motivations, audiences, and contexts of the experiments. Speaking to theorists refers to the experiments that test the predictions of the theories. Searching for facts refers to the experiments that study phenomenon of which theory might have very little to say about. These types of experiments and their replications contribute to cumulative knowledge and shape the dialogue between theorists and experimentalists. Finally, whispering in the ears of princes refers to the policy experiments and to the dialogue between policymakers and experimentalists. However, these aims have been treated rather metaphorically, and in practice, they are often intertwined.

[14] Faverau & Nagatsu (2020a) compare two historical-methodological strands of field experiments in economics: RFE's and lab-in-the-field experiements (LFE), which raise different issues regarding external validity. RFE's raise the issue of generalization whereas LFE's raise the issue of artificiality.
[15] see Ferber & Hirsc (1978) and Greenberg & Robins (1986) for social experiments; Plott (1997) for laboratory experimental testbeds; Ludwig et al. (2011) for mechanism experiments.
[16] Similar concerns about RFE's have been raised in pharmaceutical research as well (Stegenga 2016). Stegenga (2016, p. 23) shows "a wide degree of latitude in how studies are designated, executed, and analysed", which introduces biases and undermines the evidential standards of two positive results from RFE.

Theory testing and policy experiments both test hypotheses, generated by the theory or by the arguments of lawyers and lobbyists (Roth, 1986, p. 266; see also Guala 2005). The difference lies in the applicability of results: while the results of experiments *whispering in the ears of princes* are market and domain-specific, results of experiments *speaking to theorists* apply to any market (ibid.). In policy experiments, there is a controversy over a subject, theoretical arguments can be created to support both sides of the debate, and there are no previous case studies supporting any conclusions (Roth 1986, p. 262). Hence, the experiment shifts the burden of proof in these debates (Plott 1986).

Experiments that *whisper in the ears of princes* are also sometimes labelled as policy evaluations. In policy evaluations experiment is used simply as a testing procedure for the policy hypotheses (Ludwig et al. 2011, p. 3). In policy evaluation, policy is being replicated as it would be implemented at scale and tested with randomized experiment (ibid.).

Heckman has discussed the policy evaluation in more detail and distinguished between three types of questions related to policy evaluation (2005, pp. 8-9, 17). Heckman argues that the epidemiological and statistical treatment effect literature focuses only on one of the three types of policy questions. They only interest of that literature, according to Heckman, is in the impact evaluations of the historical interventions and implemented programs. Heckman argues that EBP completely ignores two other types of policy evaluation questions: taking the old and existing program to the new environment and forecasting the impacts of it; and forecasting the impacts of completely new interventions.

These tensions and debates about the randomization and policy evaluations reflect extensively discussed issues of extrapolation and external validity, discussed by many authors (see Cartwright 2007; Cook 2014; Deaton 2010a, 2010b; Deaton & Cartwright 2017; Faverau & Nagatsu 2020a, 2020b; Heckman 2019; Keane 2010a, 2010b; Leamer 2010; Ruzzene 2015; Wilcox 2016). It is not my aim to dive into this discussion here. I will follow the views of Deaton and Cartwright (2017, p. 2) that RFE's are overpromoted due to the alleged easiness of the extrapolation, but underpromoted due to the diversity of the aims that they can serve.

Deaton & Cartwright (2017) argue in their discussion about RFE's that "there is insufficient theoretical and empirical work to guide us how and for what purposes to use the findings" (p. 10). They stress the importance of recognizing what the actual hypothesis

is, and what is the purpose that evidence is supposed to serve (ibid.). They distinguish several purposes that RFE's can serve. To three inferences of them, RFE's are enough on their own (Deaton & Cartwright 2017 pp. 12-13). They are:

(i) Policy evaluation[17]

This is the case of policy evaluation, already discussed above. RFE's can be used for evaluation to demonstrate that the program achieved its goals (ibid.). But, in order to this be helpful, one would need guidelines and justifications for the usefulness of the results elsewhere (ibid.).

(ii) Counterexample to a theory

RFE's can present a counterexample to a theory or to its implications, or to confirm a prediction of a theory. "It is simply one among many possible testing procedures" (ibid.).

(iii) Estimation of population-level average

RFE's can be used to estimate the population-level average, when trial sample itself is from the same population (ibid.). However, the main threat in this kind of study would be that the individual outcomes influence the behaviour of others, which would complicate the inference from the sample to the population (ibid.).

To the following purposes, however, RFE's cannot provide a complete methodology on their own:

(iv) Simple generalization

Deaton & Cartwright (2017, pp. 10-13) discuss inference from one circumstance to another and argue that the assumption that the policy would work elsewhere, because it worked somewhere, is ultimately unwarranted assumption, which requires a justification. In some instances, justifications, such as structural reasons for the wider success of the program can be found, but a proof is nevertheless needed. They resist the idea that this extrapolation could proceed automatically. On the contrary: it would require understanding about the causal structure of the target environment and the role of support variables, which enable the functionality of the cause of interest. Researcher needs to know why things work to be able to extrapolate. However, Deaton &

[17] Deaton and Cartwright use the label of program evaluation, but the meaning is essentially the same.

Cartwright add that the discussion on external validity is usually unhelpful and does not dismiss RFE's, because results of the RFE's can be valuable in other ways.

(v) Study of local circumstances

RFE's can also be used to study the local circumstances that are conditioning the impact estimate in a given environment (Deaton & Cartwright 2017, pp. 13-14). The idea is to draw lessons about the specific population under study with the help of specified model and variation of its variables. One way to achieve this is to design the experiment to use response surface modelling like was done in negative income tax experiments.

Deaton & Cartwright also briefly discuss how average effects acquired from RFE's cannot be used on their own to understand the individual effects of the subjects (Deaton & Cartwright 2017, pp. 16-17). In contrast, other methods like good theory, analogical reasoning, process tracing, identification of mechanisms and symptoms, and sub-group analysis can be helpful to fulfil this task (Hill 1965 as cited in Deaton & Cartwright 2017, p. 17).

(vi) Extrapolation with adjustment

RFE's can also be complemented by statistical methods such as post-experimental stratification and subgroup analyses that can help to achieve extrapolation with adjustment (Deaton & Cartwright 2017, pp. 13-14). These methods would allow researcher to better predict the outcome in the target population by studying the differences and variations in the supporting variables. However, to work this method would require that these variables are equally present in the initial sample and the target population and the same equation applies to both populations. The population-level conclusions are supported only as far as we have causal and probabilistic knowledge about the properties that sample and population-level share. And this often relies on the evidence provided by other methods such as observational studies.

(vii) Scaling up

RFE's could also apply the same intervention but scaling up the experiment to a much larger area, such as to the whole country (Deaton & Cartwright p. 15-16). According to Deaton & Cartwright, this is also a case of simple extrapolation, but there are further threats with this line of thinking. The difficulty is that the effects of the scaled-up

experiment might be very different from the initial experiment, because scaling up can change the underlying causal mechanisms and levels of variables due to their interaction.

(viii) Theory building and testing

Finally, Deaton & Cartwright (2017, pp. 14-15) discuss how RCT's can be used to build and test theory. They cite the study conducted by Banerjee et al. (2015) as a leading example how one could design RCT to test a package of interventions to alleviate poverty traps and simultaneously contribute to the theory of economic development. Deaton & Cartwright (2017, pp. 14-15) notice there are a couple of ways how theory and results of RCT's can be combined.

One such a way, would include the parameter estimation in the single model experiment (ibid.). In this instance, combination of the experimental results, the baseline data from the experiment, and the structural models provided by theory, would allow them to be used in diverse policy calculations. Structural models, obviously, bring in additional assumptions concerning, for instance, functional forms and the distribution of unobservable variables, but regardless; they help to bridge the policy and theory.

Having discussed the range of possible inference that RFE's can make, I think obvious should be stated. I think it is very unlikely that one experiment would serve all these aims equally at the same time without compromising the others. I think it is equally unclear what type of inferences can serve the policymaking the most, in what circumstances, and how it really works out.[18]

[18] Here, I would like to make an additional remark. While some authors have argued that policy evaluations serve, for instance, the later stages of the policymaking, in which evidence has to be gathered on the selected policies (see Ruzzene 2015), I think it is unclear how different inferences discussed above, in fact, fit into the different kinds of policy processes. There is no universally accepted approach to policy planning. If one compares the policy process described by Ruzzene (2015), for instance, to FCC auction studies and *laboratory experimental testbeds*, discussed by Plott (1997), and the role that researcher plays in those process, the difference is striking. In these instances, researchers do not only apply different experimental methods, but are also required to make very different type of judgments in the process. While advocates of the RFE's emphasize rigorous evidence, Plott (1997, p. 631) emphasizes also judgment: "In spite of testing, rules can be incomplete and policy must sometimes be made on the spot. Decision must be made on the spot from experience and judgment".

4.2. The Finnish Basic Income Experiment

In this section, I will compare BIE Finland to both waves of RFE's. From certain aspects such as size, length, or policy aim, BIE Finland resembles more early social experiments than modern RFE's, despite it avoids the main flaws of social experiments, such as selection bias, deriving from a voluntary participation. It fulfills the main conditions of randomization, policy evaluation, and to some extent, an interest in basic behavioral relationships, characterizing early social experiments. However, BIE Finland lacks a controlled variation of the treatment variables and structural models, which were an important part of the early social experiments studying deep structural parameters, such as NIT experiments (see Heckman 2019). Also, interest in examining basic behavioural relationships is debatable. From significant parts regarding direct policy applicability, ambitious theoretical aims, or black box interventions consisting of bundle of treatments, BIE Finland does not resemble the most recent wave of RFE's. However, it fulfils three out of four main conditions of them: randomization, natural populations, and natural environment. Moreover, subjects' awareness of the experiment can possibly introduce biases, especially in the treatment group, which decreases the usefulness of the experiment for theory testing.

The aim of this section is to help the reader to understand to what extent BIE Finland can be interpreted as speaking to theorists and guiding policymaking. The comparison to the alleged strengths of the modern RFE's and to earlier social experiments highlights the ambiguities of the experimental design that reflect various value choices.

Based on the previous section and the possible aims of the RFE's distinguished by Deaton & Cartwright (2017), I will offer an interpretation of the type of inference in BIE Finland. I will suggest BIE Finland can be interpreted as (1) a simple testing procedure of the claim generated by the policymakers or theorists, a (2) policy evaluation or (3) an attempt to extrapolate.

BIE Finland randomly assigned two thousand unemployed people aged 25-58 to receive monthly allowance of 560 euros (Kangas et al. 2019, pp. 8-9). It aimed at replacing the main parts of the existing social benefit system of Finland, such as unemployment and housing assistances. The effects of the basic income were studied for a two-year period, 2017-2018, in terms of the labour market supply and the recipient's welfare, while

remaining unemployed people (n=173 000) formed a control group, to which the results were then compared.

BIE Finland has largely been celebrated by the public as a uniquely extensive and first nation-wide social policy experiment, and it has yielded global media attention (see Kangas et al. 2019; Yle 12.5.2017). It has also been severely criticized (see Kela 2016, pp 59-60), which makes it a good study to illustrate controversies and ambiguities with the EBP.

Let me start with its resemblance to the social experiments. Social experiments aim at evaluating the aggregate economic and social effects of the experimental treatment that is a policy by a government agency (Ferber & Hirsch's 1978 as cited in Harrison & List, 2004, p. 1036). They can also be defined explicitly as tests of social welfare policy (Greenberg & Robins, 1986, p. 341-342). Overall, social experiments have often targeted at disadvantaged people or families, such as public assistance recipients, low-income families, unemployed, and youth (Greenberg & Shroder 2004, p. 159-160).

Other authors have also emphasized the importance of randomization as the defining aspect of social experiments. Ferber and Hirsch define Social Experiments as "rigorously defined study incorporating experimental treatments applied over a period of time to statistically randomly selected parts of a population" (Ferber & Hirsch, 1978, pp. 1380-1381). Also, Greenberg & Robins (1986, p. 341-342) list randomization as a defining feature of social experiments, distinguishing them from mere demonstration projects. More recently, Greenberg & Shroder (2004 as cited in Levitt & List, 2009, p. 4) have highlighted four criteria for Social Experiments: (i) random assignment, (ii) policy intervention, (iii) follow-up data collection, and (iv) evaluation.

BIE Finland seems to resemble these aspects of social experiments well. It studied both social and economic effects of the basic income in terms of the labour market supply and the recipients' welfare. The treatment group of 2000 people was randomly selected from the people who were, according to the register data of the Social Insurance Institution (Kela), unemployed in November 2017. Unemployed people that were not selected to the treatment group formed the control group. Follow-up data on subjects' welfare was gathered via surveys.

Early social experiments (1962-74) were "elaborate, lengthy, costly" social experiments that tested "fundamental changes in social policy" (Greenberg & Robins 1986, as cited in

Greenberg & Shroder, 1999, p. 158). Over 80 percent of all these social experiments tested completely new programs (ibid., p. 162). According to Levitt & List, these large and costly social experiments mirrored the social policy debates about the welfare system in the US in 1960's (Levitt & List, 2009, p. 5). Also, de Souza Leao & Eyal (2019, p. 399) have noticed that the first wave of RFE's consisting in social experiments were typically longer and bigger in geographical scope, and they evaluated "whole delivery systems".

BIE Finland was also relative lengthy as the subjects were studied over two-year period. Good and comprehensive register data of the Finnish administration also allows a follow-up data collection for much longer time. BIE Finland was also a nation-wide experiment and the subjects were randomly allocated to the treatment group from different parts of Finland (Kangas 2016; Kangas et al. 2019, p. 7). The Finnish debate on the basic income also represents it as somewhat fundamental reform of the social benefit system, despite critics argue that the experiment was much more modest in this regard. The budget for the experiment including the administrative costs was 20 million euros (Kela 2016, p. 60)

In the early phases, social experiments studied "deep structural parameters" and "basic behavioral relationships" that were used to evaluate a broad range of social policies (Levitt & List 2009, p. 6). According to Levitt and List, the most optimistic people thought that even policies, which were not even implemented, could have been evaluated on this basis (ibid.).

Ultimately, BIE Finland studied the effects of the basic income on the employment and income (Hämäläinen et al. 2019, p. 2; Kangas et al. 2019, p. 9). Thus, it was the relationship between the unconditional cash transfer and the labor supply that was the main research interest. This relationship has been in the interest of economists for a long time and somewhat similar studies and experiments have been conducted already around the world (for summary see Banerjee & Duflo 2019, pp. 277-322). The relationship of the cash transfers and the employment effect is connected to potential disincentives for work that economists call as *income effect* and *substitution effect* (ibid., p. 290). According to the former, additional money is less needed when social welfare assistance is being received, and it therefore disincentivizes the work (ibid.). According to the latter, working is less valuable if social welfare assistance is being received, because the more I earn from the work, less is being received as welfare assistance (ibid.).

BIE Finland aimed at creating a system, in which the *substitution effect* would have been diminished by removing some disincentives for work. The idea of the unconditional basic income was that person would not lose anything from the basic income, even if they had accepted work and earned more (Kela 2016, p. 8; Kangas et al. 2019, p. 7). In practice, the implementable basic income model would have had a progressive tax model (Kela 2016, p. 61), meaning that substitution effect would not have been removed entirely. However, in the experiment the basic income was defined as a tax-exempt income (ibid.). This tax-exempt was justified by the power analyses and, according to the research group, it was necessary to keep the economic incentives sufficiently high in order to observe the potential effect with the given sample size (ibid., pp. 61-62). As a result, in the experiment the marginal tax rate for work was decreased 30 percentage points at maximum (Hämäläinen et al. 2019, pp. 1, 8).

Can BIE Finland, thus, be interpreted as studying the size of the income effect of the given basic income in the Finnish society from a theoretical point of view? No, not really. This is complicated by the fact that the experimental design also included components that simultaneously disincentivized the work, such as removing waiting periods and responsibilities to attain activation measures (Hämäläinen et al. 2019, pp. 1, 6-7). Thus, separating the effects of these two mechanism, incentives and disincentives, was not possible (see Kangas et al. 2019, p. 31). However, the research group summarized that the basic income model allows them to study the effect of financial incentives to people's behavior and how bureaucracy traps affect behavior (Kangas 2016; Kela 2016, pp. 60-61).

Let me discuss two other important features of the experiment: the model choice of the experiment and the degree of variation in the treatment variables.

BIE Finland studied only a single treatment of basic income with the support level of 560 euros, although another experiment with varying level of support was recommended by the research group (Kela 2016). In fact, the Finnish constitution ruled out the possibility of varying the level of support for other recipients on the grounds of equality (Kela 2016, p. 60).

Compare this with the Negative Income Tax (NIT) experiments, an example of early social experiments. The New Jersey Income Maintenance Experiment is often considered as a spark for a wider interest in experimental methods in economics (Levitt & List 2009,

pp. 2-5). The New Jersey Income Maintenance Experiment was proposed to a governmental agency by Heather Ross and sponsored by the Office of Economic Opportunity (OEO) (ibid.). This resulted experiments in five urban communities in New Jersey and Pennsylvania, with the price tag of over 30 million dollars in today's currency (ibid.).

The New Jersey Income Maintenance Experiment aimed at evaluating the economic and social effects of various experimental treatments (Ferber & Hirsch, 1978, p. 1380-1382). The experimental treatment was determined by two variables: the support level and the tax rate. Both variables were varied to evaluate their combined effects on the work behaviour of the beneficiaries. The subject pool consisted in the low-income families, whose head of family was between 21 and 58 years at the time of experiment (Ferber & Hirsch, 1978, p. 1380-1382). In total, around 1300 male subjects took part in the experiment. The subjects were randomly selected. According to Ferber & Hirsch, the need for the experiment derived from the difficulties of theory to predict the magnitude of the response of income maintenance**.** Additionally, previous efforts to make an estimate yielded different results, so that there were uncertainties regarding both the effects of the program and the price of it.

So, the NIT experiments studied the labour supply effect, but with the help of structural model, and by systematically varying both the tax level and the level of supply (Card, DellaVigna, & Malmendier 2011, pp. 54-55). The largest NIT experiment, Seattle-Denver Income Maintenance Experiments, consisted in as many as 58 distinct treatment groups with their own level of support and tax rate (ibid.). Moreover, the inferences could be done only with the help of the specified structural response model (ibid).

This highlights how early social experiments were testing new programs by deliberate variation of the policies (see also Greenberg & Shroder pp. 159-160; Levitt & List, 2009, pp. 2-5). This means that the treatment, a policy by a government agency, was varied to evaluate program's effects (Harrison & List, 2004, p. 1036). Structural model can increase the ability to extrapolate to the target environment (see Ruzzene 2015).

If BIE Finland is compared with the NIT experiments, it seems that no explicit structural model was being tested in BIE Finland. Neither was extensive variation of the parameters introduced in BIE Finland. It seems that BIE Finland resembles what Card et al. (2011, pp. 42) call "descriptive studies that lack any formal model". According to them, most

field experiments have always been descriptive, although theoretical models have become more popular lately (ibid., pp. 46-48).

However, the experimental design of BIE Finland allows it to be used for the impact evaluation of multiple different components of the study (Hämäläinen et al. 2019, p. 1). The register data allows to identify different sub-groups[19], such as those who applied for unemployment benefits and those who did not, that can be used to analyse the effects of unconditionality. Different sub-groups regarding the marginal tax rate for the work can be also distinguished on the basis of family structure and living costs.

The lack of theory in the design and the nature of the BIE Finland as a descriptive study leads us to interpret that it did not aim at studying deep structural parameters and basic behavioral relationships in the same way as early social experiments. Rather it should be understood as a testing procedure of a single treatment that tests the effects of certain type of basic income market- and domain-specifically in the Finnish context.

Card et al. discuss the optimal level of theory in the experimental design and notice that many have come to favour the precise estimation of a small number of treatment effects (Hausman & Wise 1985, p. 188; as cited in Card et al. 2011, p. 56). Too many treatment groups have, for instance, jeopardized the conclusions because of the too small sample sizes (Card et al. 2011, pp. 55-56). This seems to be the case in BIE Finland also, as there was already an issue with the small sample size (see Kela 2016, p. 61) as well as with legislative issues concerning the variation. But there is no real consensus regarding the optimal level of theory and modelling (Card et al. 2011, pp. 55-56).

However, there have been some well-recognized epistemic risks in the methodology of the social experiments. For instance, the behavioral responses of the subjects might be influenced by the experimental methodology, which brings the danger of biased results (Rawlings 2005, p. 195). If participants apply to the program voluntarily there is a possibility of selection bias. Moreover, either the treatment or control group might adjust their behavior in response to the evaluation itself, which are known as Hawthorne effect (treatment group) or John Henry effect (control group). The attrition bias might also arise as the subjects might voluntarily leave the experiment, either from treatment or control

[19] Subgroup analysis was done, but groups were too small to draw reliable inferences about them and the inference suffered from the problem of multiple hypothesis testing (Hämäläinen et al, pp. 18-21). In this regard, BIE Finland seem to rely more on straightforward randomization, and hypothesis-testing (of the primary variable), rather than on art of measuring (see Hämäläinen et al. 2019, p. 16; Wilcox 2016).

group. The subjects in the control group might also seek for the substitutes for the program, which is called as substitution bias.

In the BIE Finland, subjects were aware of the experiment, which makes John Henry and Hawthorne effects possible. However, as the subjects were assigned to the treatment group involuntarily, no selection bias or attrition bias are possible. It is also hard to imagine how the control group could seek substitutes for the basic income. Perhaps illegal work can have caused substitution effect, which can potentially bias the findings, if the recipients did not report their working hours honestly.

Levitt & List (2009, pp. 16-19) contrast the most recent wave of field experiments with social experiments and argue that the latter can avoid many shortcomings and well-recognized biases of social experiments. They argue that from the beginning of the 21st century, economists have been increasingly using field experiments to study a wide range of economic phenomena, which marks the most recent wave of experimentation. Randomization, natural populations, subject's natural environment, and unawareness of being part of the experiment all characterize the aspects of the most recent field experiments (Levitt & List 2009, p. 19).

BIE Finland meets three out of four these criteria. Firstly, the subjects were randomly assigned. Secondly, the experiment also took place in the natural environment of the Finnish society. Thirdly, it also studied natural population of the Finnish unemployment benefit recipients nation-widely. Even though this population does not represent the characteristics of all Finnish people and many distinct groups were excluded, the subject population can be interpreted as a natural population to some extent (Hämäläinen et al. 2019, pp. 10-12, 14). However, it is notable that studying only one group could limit the conclusions, if the target group was not otherwise justified (Harrison & List 2004, pp. 1012-1013). In BIE Finland, this target group was justified by practical concerns of making sampling speedy and effective as well as keeping the administrative costs of model implementation lower (Kela 2016, p. 61). The research group is explicit that sample did not form representative sample of Finnish population, which makes extrapolation hard (Hämäläinen et al. 2019, p.12). However, the choice of the treatment group was justified by their relevance as employment would increase their own wellbeing (ibid.). Fourthly, as it was mentioned, the subjects who received the basic income were

obviously aware of the arrangement. However, it is also possible that the people in the control group did not know that they were studied too.

More specifically, the biases that according to Levitt & List (2009) can be avoided via the most recent wave of RFE's include randomization bias, attrition bias, Hawthorne effect, and substitution bias. Much of these advantages seem to relate particularly to natural field experiments, in which subjects are unaware of the experiment (see Levitt & List 2009, p. 26). For this reason, these claims do not hold very well with the BIE Finland. As it was mentioned, in the BIE Finland the treatment group was aware of the experiment, so Hawthorne effect or substitution bias could not have been avoided. Only attrition bias was not possible in the BIE Finland as the subjects could not have been withdrawn from the experiment. Neither was there significant randomization bias regarding observables (Kangas et al. 2019, p. 11). Despite one variable was imbalanced between treatment and control groups, it did not significantly affect the conclusions as it was controlled in the statistical analysis (ibid.).

So, as it is clear, BIE Finland is not a natural experiment that would meet all the conditions of the natural and realist environment (see Harrison & List 2004). In terms of Harrison & List (ibid.), BIE Finland is a framed field experiment that differs from natural experiments mainly because subjects are aware of the experiment (see Harrison & List 2004, pp. 1012-1014). Otherwise, regarding the subject pool, commodity, tasks and trading rules, or the context and information sets of the experiment, framed field experiments include more realistic aspects than laboratory experiments (ibid.).

Even though natural experiments seem to be the ideal of the most recent wave of economics experiments, framed field experiments do still play a major role in the recent wave of experimentation (see Levitt & List 2009, pp. 24-26). They have been applied to inform policymaking and to contribute to the theoretical literature (ibid). List (2011, pp. 5-6) makes clear, even framed field experiments can look very different from each other. They can include social experiments from 1960's to 2000s, randomized controlled trials conducted in developing countries, or experiments that test theories (ibid.).

These differences in the methodology of framed field experiments is also important for the present discussion. According to Levitt and List (2009, p. 24), some of these studies deploy treatments in a more straightforward manner so that treatment consists in several variables, but which are all directly linked with the policy alternatives. According to

Harrison and List, the nature of the social experiments also changed over time so that experiments could not be seen only as systematic variation of certain policy (Harrison & List, p. 1037). As experiments were designed to test incremental changes to existing programs, experiments typically tested various packages of services and incentives all at once (Levitt & List 2009, p. 6). This resulted in "black box" social experiments[20] that were testing a bundle of variables, all potentially influencing behavior (ibid.).

However, in the BIE Finland there are no confounding variables or packages of services being tested simultaneously. In this regard the policy tools available to achieve the aims associated with the basic income are limited to economic incentives only. Perhaps the results can be compared with the data regarding the services and activation measures for the unemployed.

So far, I have discussed the characteristics of the BIE Finland and compared them with social experiments, and the most recent wave of RFE's. This comparison highlights how framed field experiments can serve theory testing and policymaking in varying degrees. A summary is provided below in table 1.

[20] This type of approach can be seen particularly clearly in the most recent wave of experiments in development economics. Consider for instance Duflo et al. (2006), who compared three school-related HIV/AIDS interventions in Kenya and their impact on teenage childbearing: training teachers, encouraging students to debate the role of condoms, and reducing the cost of education.

Characteristics	1st Wave of RFE's	BIE Finland	2nd wave of RFE's
Size and length	*elaborate, lengthy, costly*	*elaborate, lengthy, costly*	*short-term*
Policy aim	*yes: policy evaluation, model building and prediction*	*yes: policy evaluation, population-level average treatment effect*	*yes: policy evaluation of direct policy alternatives*
Randomization	*yes*	*yes (individual level)*	*yes*
Natural populations	*yes*	*yes* (no representative sample of Finnish population, but of unemployment benefit recipients)*	*yes*
Natural environment	*yes*	*yes*	*yes*
Basic behavioral relationships	*yes*	*no*	*no/yes*
Number of treatment arms	*many: as many as 58 treatment arms*	*a single treatment with no controlled variation*	*a precise estimation of a few*
Controlled variation of the treatment	*yes*	*no*	*yes*
Structural model and parameter estimation	*yes*	*no*	*no*
Ex ante theory (i.e. risk attitudes, time preferences)	*no*	*no*	*yes*
Ambitious theoretical goals	*no*	*no*	*yes*
Direct policy applicability	*no* (but extrapolation within the model choice)*	*no*	*yes*
"Black box" interventions	*no*	*no*	*yes*
Ex post theory	*yes*	*no* (but follow-up data gathering with the register data, although no extrapolation possible because of sample size)*	*no/yes*
Possible Biases			
Voluntary partcipation	*yes*	*no*	*yes/no*
Selection bias	*yes*	*no*	*no*
Subjects' awareness	*yes*	*yes*	*no*
Hawthorne effect & John Henry	*yes*	*yes*	*no*
Attrition bias	*yes*	*no*	*no*
Substitution bias	*yes*	*yes/no*	*no*
Randomization bias	*yes*	*yes (no significant effect)*	*no*

But, where does this leave BIE Finland regarding its epistemic aims? Let us recall the aims of RFE's distinguished by Deaton & Cartwright and discussed in the previous section. I suggest that BIE Finland can be interpreted as (1) a simple testing procedure of the claim generated by the policymakers or theorists, a (2) policy evaluation or (3) an attempt to extrapolate.

As a testing procedure for the theoretical claim, however, the experimental design was suboptimal, because the subjects in the treatment group were aware of the experiment. Also, the lack of theoretical model and controlled variation can make it harder to apply the results later for both theory-building and future policy evaluation. Furthermore, because the level of incentives was determined by the allocated budget, and not by the previous knowledge about the effectiveness of the monetary incentives, it raises a doubt regarding the experiments' possible contribution to theory.

As a policy evaluation, BIE was not an evaluation of a real implementable program, but perhaps some parts of it, because the study did not include a representative sample of the Finnish population or involve all the same features. Recall that experiment had different tax model than implementable model of basic income would have had. Policy evaluation did neither involve bundles of services or direct policy alternatives generated by theorists or policymakers, which makes the experiment, perhaps, resemble more like conventional theory experiment than policy evaluation. Even though, subgroup analysis would be helpful to some degree, in designing new policies, it does not fulfil the evidential standards of the EBP.

This brings us to the third interpretation that BIE Finland was an attempt to extrapolate the effect of the monetary incentive from the experimental target group to the nation-wide population-level average, conditional on the characteristics of the target group (unemployment benefit recipients). This seems like a technically accurate description of the inference, but I think it leaves open both questions: what value the experiment bears on the theoretical literature or the future policy design. However, conditionality on the given sample seems to be something that was not planned. It seems that the aim was initially to extrapolate to the national level so that the results could be generalised to the entire population (Kangas 2016, Valtioneuvoston kanslia 2016, p. 12).

Therefore, it is not entirely clear what the epistemic aim of the BIE Finland was, or should have been. Because of this unclarity, it is not obvious whether all respective features,

important for the two waves of RFE's, were realized or not in the experimental setting. But I think it is safe to conclude that we can at least suspect suspect that the experiment was suboptimal in more than one way as we can see potential improvements regarding all potential aims of (1) simple testing procedure, (2) policy evaluation, and (3) an attempt to extrapolate. Whether this has consequences for objectivity is another matter.

5. TRAINED JUDGMENT AND VALUES IN RFE's

5.1 Beyond mechanical objectivity

In this chapter I will explore what implications a more realistic understanding of the science-policy interaction will bear on objectivity of the RFE's. Co-creation as manifested in conducting of RFE's in policymaking, raise the need to understand the implications for the normative guidelines of the scientific process. As I argue, co-creation presents both epistemic risks and opportunities for the research. I will discuss how scientific process is altered and influenced by the co-creation of the RFE's in BIE Finland, what role non-epistemic values play in the process, and elaborate the potential consequences for objectivity of the research. I will argue that uncertainty and risk of error ultimately urge researcher to invoke a human judgment that goes beyond mere mechanical objectivity.

This chapter is divided into two mains sections. After introductory section (5.1), next section (5.2) deals with EBP and the value-free ideal. Following de Souza Leao & Eyal (2019), I will demonstrate how EBP relies on autonomous and authoritative view of science**.** This view has historically coincided with the value-free idea to a large degree (Douglas 2009, pp.7, 44-65) and it seems to be the case with EBP as well. The value-free ideal gained momentum in 1960's as an attempt to protect science from the pressures of the society (Douglas 2009, pp. 60-65). It is still a popular view despite its long-lasting criticism that has often been ignored or marginalized (ibid). I will follow Khosrowi (2018) in arguing that EBP is committed to the value-free ideal. According to the value-free ideal, good science is to be made without any political or ethical considerations interfering scientific reasoning and practices (Douglas 2009, p. 45; Reiss & Sprenger 2017).

After having established the connection between the value-free ideal and EBP, in the following section (5.3), I will argue that viewing science as authoritative and value-free endeavour present epistemic risks and opportunities for the EBP when scientific methods are being co-created by researchers and policymakers, as manifested in BIE Finland. Following Risjord (2014), Douglas (2000, 2009), and Longino (1990), I will show that ethical and political values can, and often become, as constitutive values of the research. Often this will happen via *inductive risk* (Douglas 2000, 2009; Reiss and Sprenger 2017; Risjord 2014). Inductive risk arises whenever there are two potential errors, such as

overestimating and underestimating the treatment effect, which cannot be reduced simultaneously (Risjord 2014).

Based on the combination of the theoretical literature and the BIE Finland, I will illustrate how political and ethical values can become constitutive values of the research through technical decisions of the experimental design (5.3.1), theoretical content (5.3.2), or accepting and rejecting hypotheses (5.3.3).

All three routes present challenge to the value-free ideal on which EBP relies. They also demonstrate three different scenarios of inductive risk when scientists (1) make decisions about the experimental design, (2) select appropriate theory, and (3) interpret the results, illustrating how researchers are forced to face the risk of error and uncertainty when conducting RFE's for policy purposes. They also exemplify points of decision where human judgment is needed and in which both epistemic and non-epistemic values can influence the decisions of the researcher.

The need for human judgment, and related type of individualistic process objectivity, has also been elaborated by several authors in economics and other disciplines as well (see Galison 2015, p. 58; Reiss 2014, pp. 138-140). Gaston (2015, p. 58) contrasts the expert qualifications of the observer with strict inherent procedures and protocols, highlighting a long and careful training and scientists' capacities to re-identify patterns, eliminate artefacts, and categorize the world. For Reiss (2014, pp. 138-140), *considered judgment* about scientific hypothesis would require scientists to think all the evidence relevant to the assessment of hypothesis. This would include judgments about relevance, the quality of evidence, and the need for additional evidence (ibid.). Researchers would have to balance and weight the pieces of evidence, and to decide how much evidence is needed to accept or reject the hypothesis, and at what costs it could be pursued (ibid.). These judgments rarely follow any strict rules but are dependent on the expertise of the researcher, which is why these judgments are sometimes perceived as more subjective than mechanical objectivity (ibid.).

Reiss seems to consider a considered judgment especially in terms of confirmation and rejection of the hypotheses. While this also seems to be the case with BIE Finland, I will expand on the ideas of considered judgment and trained judgment beyond mere confirmation of the hypotheses to include other stages of the research as well, including theory choice, experimental design, and interpretation of the results. As I will argue,

trained judgment and considered judgment should be recognized and discussed more explicitly by the EBP. They show a wide array of routes how human judgment can evoke epistemic risks for the research and cause trade-offs between epistemic and non-epistemic values. As I will argue the magnitude of the epistemic risk caused by non-epistemic values also varies, and in some instance non-epistemic values also provide opportunities to increase objectivity of the research as they can help researchers to decrease the risk of error.

Notice also that the criticism evoked against 1980's applied econometrics was directed precisely towards the inadequate, or completely absent, judgments concerning the quality of the causal evidence. Also, consider the fourth driver of the experimental turn, the demand that experimental data should be considered in tandem with economic theory (virtuous cycle) (Svorencik 2015, pp. 5, 238). This is a demand about the relevance of the experimental data, invoking a considered judgment. Angrist & Pischke make a similar point in the context of credibility revolution when they argue that good research designs complement good research questions (2010, p. 25).

Yet, it is unclear to what degree inductive risk has consequences for objectivity (see Douglas 2016, Elliott and Richardson; Stegenga 2017, p. 19). Some authors have suggested that demarcating warranted and unwarranted instances of non-epistemic values, requires a reference to the conventional standards of the research community (Wilholt 2009 as cited in Stegenga 2017, p. 28). Others have pointed out how these conventional guidelines can sometimes be too easy to satisfy any reasonable evaluation criteria, even when they require evidence from more than one RFEs, as is the case with pharmaceutical research (Stegenga 2017, pp. 23-28). According to Stegenga (ibid., p. 28), deviation from community standards can be justified on both epistemic and non-epistemic grounds.

This challenges EBP to develop more explicit account of the risks of error throughout the research process, going beyond the value-free ideal, and involving elaboration of potential trade-offs between epistemic and non-epistemic values and their consequences for objectivity.

5.2 The value-free ideal

In this section, I will show that EBP is committed to the value-free ideal, impartiality, and neutrality. I will also illustrate ambiguities regarding their views concerning the role of values in the scientific process that can potentially conflict with impartiality.

The value-free ideal has been an influential account of objectivity and the majority of the modern discussion about objectivity is centred around the topic (Koskinen 2018, p. 4; Tsou et al 2015, p. 2). It is a long-lasting tradition in the philosophy of social sciences to consider normative value commitments as a potential source of bias undermining scientific objectivity (Reiss & Sprenger 2017). The value-free ideal states that scientific endeavour should therefore be as free as possible from social and ethical values that do not belong to the core of the scientific community (Douglas 2009, p. 45).

There are two interpretations of the value-free ideal that are important for the subject of this thesis: impartiality and neutrality (see Lacey as cited in Reiss & Sprenger 2017; Risjord 2014, p. 20). As I will argue later, both are implicit to some degree for EBP. According to neutrality, scientific evidence is descriptive in the sense it tells how things are, not the way how they should be (ibid.). According to impartiality, conclusions of the research are to be drawn without ethical and political values directly influencing them, and by their contribution to scientific values (ibid.). However, closer inspection also reveals controversies and ambiguities regarding how these should be interpreted in the context of EBP.

Lacey (as cited in Reiss and Sprenger 2017) distinguishes autonomy as third component of the value-free ideal, implying that the scientific agenda is determined by the scientific and not political and ethical interests. I will not discuss autonomy as it has been criticized and rejected elsewhere (see Douglas 2009, p. 17; Reiss and Sprenger 2017). It is also less relevant for EBP given their implicit aim of producing evidence for policy purposes. Impartiality is also logically prior to neutrality and autonomy (Lacey as cited in Douglas 2009, p. 17). According to Risjord (2014, p. 20), impartiality and neutrality present two alternative ways in which ethical values can become constitutive of the research, thus presenting two debatable issues for the value-free ideal. Either ethical values can influence the process of justification and confirmation of the hypotheses or they can become as a part of the theory (Risjord 2014 pp. 20, 30). Let me discuss them in turn, starting with the former.

5.2.1 Impartiality

The value-free ideal can be divided into a strong and a moderate thesis depending on how extensively and what type of values are prohibited (Risjord 2014, pp. 17-20). The Strong Thesis of Value Freedom holds that all values are prohibited in scientific research (Risjord 2014, p. 17). This position is too demanding and cannot be maintained (Douglas 2009, pp. 90-91; Risjord 2014, pp. 17-19). Without any values guiding the practice, researchers could not make any methodological choices (ibid.).

The moderate thesis holds that only values internal for science can be fundamental and constitutive part of the research (Risjord 2014, p. 18-19). According to Risjord, the moderate thesis of value-free ideal distinguishes between *epistemic* and *non-epistemic values.* Many philosophers have recognized them as distinct value categories and have allowed only *epistemic values. Epistemic values* refer to the internal values of the scientific community that contribute to good science and are therefore acceptable, whereas *non-epistemic values* include all wider political, ethical or cultural values, which are forbidden (ibid.; Douglas 2009, pp. 89 - 90). Examples of the epistemic values include predictive accuracy, explanatory power, scope, and simplicity, which are clearly related to scientific reasoning (Douglas 2009, pp. 89-90). Non-epistemic values cover for instance "concern for human life, reduction of suffering, political freedoms, and social mores" (ibid.).

According to the value free-ideal, these two types of values have different consequences for objectivity: "epistemic values are not threatening to objectivity, while moral and political (non-epistemic) values can be potentially troublesome" (Risjord 2014, p. 30). Science that avoids moral and political values directly interfering preference for one conclusion over another or the methodological choices, is thus labelled as *impartial* research (Risjord 2014, p. 20).

The notion of impartiality is often supplemented by the observation that not all non-epistemic values present a threat for the scientific reasoning. The value-free ideal does not prohibit all non-epistemic values to influence scientific process (Douglas 2000, pp. 563-564; Longino 1990, pp. 83-86; Risjord 2014, p. 19). N*on-epistemic values* can shape the scientific activity, but they cannot be a necessary condition for the activity (Risjord 2014, pp. 18-19). There is also a distinction between c*onstitutive values* that are necessary for the activity in the sense that activity could not go on without commitment to them and

contextual values (ibid.). C*ontextual values* are part of the environment. C*onstitutive values* "shape the activity from the inside" (p. 18), whereas *contextual values* might "shape the activity, but they are not necessary to conducting it" (ibid., pp. 18-19). Therefore, according to the moderate thesis of value-freedom, non-epistemic values should be contextual and not constitutive of the scientific practice (ibid.).

Values that are counted as constitutive, and thus necessary for the activity, reflect our understanding and definitions of the essentials of the science. Process-wise, the scientific practice can be divided into four core stages: (a) selection of a scientific research problem, (b) gathering of evidence, (c) the acceptance of a theory or hypothesis based on the evidence, and (d) proliferation and application of results (Weber 1917 [1988] in Reiss & Sprenger 2017). In addition to research ethics, non-epistemic values are generally allowed to play a direct role in (a) the selection of a research problem and (d) the application of the technologies (Douglas 2000, pp. 563-564; Longino 1990, pp. 83-86).

This can be illustrated with the BIE Finland. Firstly, the interests of the politicians and policymakers obviously influenced the selection of a problem through the legislative bill and the terms of the reference of the research. It was in the government's main interest to study specifically the employment effect of the basic income (Kela 2016, p. 58). Moreover, the legislative bill states, for instance, that improving the employment situation of students or persons receiving old-age pensions was not the objective of the experiment, justifying their exclusion from the testing group (ibid.). Secondly, The Finnish Constitution ruled out the possibility of varying the level of welfare assistance based on equality consideration, which was a matter of research ethics (ibid., p. 60). Thirdly, despite the wishes of the research group (ibid., p. 62), it is yet unclear if the results will be applied and proliferated either as policy reform or as follow-up study in future**.**

Risjord (2014, p. 19) notices how political values quite often determine the research programs that get funded, but it is not necessarily a threat to objectivity. He suggests "funding science is a little bit like shining a flashlight into the dark: Interests determine the direction of the beam, but not what we see when we look" (ibid.). In his classical writing, Nagel (1961 pp. 485 - 486) also argues that the selection of the research problem, even though it is influenced by value considerations is not an issue for the objectivity of science as it applies equally to both natural and social sciences alike.

Most philosophers therefore accept that ethical and political values, individual preferences, and socio-economic-cultural landscape, can influence the selection of the problem and, hence, the first phase does not belong to the core of the value-free science (Reiss & Sprenger 2017). No interests of individual researchers, funding parties, or wider social landscape pose a threat to objectivity of research when it comes to the selection of problem (Reiss & Sprenger 2017).

According to the standard interpretation of the value-free ideal, political values determining the research choice would be as legitimate as any other research choice. Thus, they are not threatening impartiality. Science is an open and competitive endeavour, and there are no pre-set rules governing what phenomenon should be studied. While there might be better or worse research topics, discriminating between the research interests is very difficult to justify.

However, I think Risjord's analogy is potentially misleading. Whether research funding is directing the beam of flashlight into the darkness without determining what will be observed is a questionable proposition. This is because political interests and financial values influencing the selection of a problem in BIE Finland, also had consequences for gathering of the evidence via technical design decisions thus, affecting the type of inference that could be drawn in the experiment. Whether this illustrates a violation of the value-free ideal or shows inadequacy of the ideal, will be discussed later in section 5.3.1. Before that let me discuss how EBP and the ideal of impartiality are related.

According to Khosrowi (2018, p. 5-6) it is hard to say how EBP relates to the values as EBP literature "rarely comment on value-related issues". However, he argues that EBP relies on the idea that some level of division of labour is possible between normativity and factuality. Moral values, according to EBP, are prohibited and only values that are internal to science are allowed for scientists. Thus, the researchers are resolving factual matters independently of political disputes. They only offer hypothetical imperatives: normative claims that are conditional on the specific political aim or values.

I agree with Khosrowi's analysis that EBP preserves the distinction between epistemic and non-epistemic values and presupposes the idea of impartial research. This is no surprise given how influential the value-free ideal has been in the past decades (see Douglas 2009). According to Douglas, there is a longstanding prohibition to use values in the place of evidence, which is often and unfortunately expanded to more overarching

ideal of value-freedom (Douglas 2004, pp. 459-460). She has also noticed how mechanical objectivity often conflates various aspects of objectivity: such as ideas about social processes eliminating biases, ideas concerning restrictions of using values, and the idea that personal interpretations can interfere reliable results.

Khosrowi (2018, pp. 3-4) summarizes the main epistemic values of the EBP as methodological rigour, unbiasedness, precision, and ability to obtain causal conclusions. All these values seem to be closely connected with the methodological choices such as randomization and a lack of theoretical assumptions, which support conclusions about policy effectiveness.

Overall, epistemic values of the EBP seem to connect very closely with methodological preferences that have been largely discussed in the literature. Deaton & Cartwright (2017) discuss unbiasedness and precision in detail but notice that precision does not follow automatically. Natural environment, realism, and level of control via randomization, are often emphasized (see Harrison & List 2004, p.1011; Levitt & List 2009, p.2; Morton & Williams 2010, p. 46). Possibility to use statistical tools both in the design and analysis phase has also been recognized as an essential feature, making RFE's better than other methods (Morton & Williams 2010, pp. 93-94). The lack of theoretical assumptions is also important, but also a debatable virtue (see Deaton & Cartwright 2017, pp. 2-3; Harrison 2013).

De Souza Leao and Eyal (2019, pp. 392-397, 409) have also discussed the value-basis of the most recent wave of RFE's from a sociological point of view. According to them, the second wave of RFE's has been successful partly due to the common set of ascetic values that are uniting different fields evolving around development aid. They refer particularly to trust in numbers, leverage, and libertarian paternalism as a common set of values (pp. 409-412).

According to them, trust in numbers is manifested in the EBP's emphasis on measurement and constant feedback loop that does not account for already accumulated theory, experience, and expertise of the development community (ibid.). Leverage is realized in the randomistas' emphasis on small interventions and nudges as a part of the broader portfolio approach. After investigating what are the most effective policies, one can then leverage and scale up the most successful ones for other institutions and governments. Libertarian paternalism connects with behavioural economics and the idea of nudging. It

presents randomistas as choice architects who steer the choices of the individuals for their own good without violating individual autonomy and using coercive power.

While their discussion is insightful and remarkable regarding the rise of the second wave of RFE's, it does not tell too much about epistemic values underpinning the EBP. I think they are also right in pointing towards the limited set of values underpinning the EBP. Trust in numbers, mechanical objectivity, and preference on small-scale studies (p. 404) are all clearly related to scientific reasoning, but they are too superficial to properly understand the value-basis and choices of the researchers.

I have here discussed value-free ideal in terms of impartiality. It seems clear that EBP is committed to the ideal of impartial research that coincides closely with mechanical objectivity. EBP is clear on methodological rationale of the RFE's and epistemic values that underpin their methodological choices justifying RFE's. Simultaneously, their discussion about the role of non-epistemic values is largely missing and they presuppose some level of a distinction between factuality and normativity.

5.2.2 Neutrality

Another way to consider the value-free ideal, in addition to impartiality, is to consider the type of evidence that is being produced (Risjord 2014, pp. 20, 24). Science can be said to be *neutral* if results do not include 'oughts', normative ideals, or binding norms. So do RFE's provide results that include oughts, norms, or ideals?

According to Khosrowi (2018, p. 6), EBP is committed to some type of value-neutrality as illustrated by its commitment to hypothetical imperatives, division of labor between factuality and normativity, and separating the roles of researcher and policymaker in this regard. He argues that EBP does not present "unconditional normative claims regarding desirability of social outcomes" (ibid.). However, Khosrowi leaves it open whether the results can be interpreted as normative. If they can, then they will be conditional on the value premise, but they do not themselves endorse the value premise of the policymaker.

Niiniluoto (1993) has discussed the logical structure of technical norms that are factual statements about means-ends relations, which are essential for applied and design sciences. According to him, technical norms are normative, but their source of normativity can vary. Normativity of the EBP and the results of RFE's, derive from policy

effectiveness, which is a specific type of effectiveness: effectiveness on average (Khosrowi 2018, p. 23). This can be compared with other potential interpretations of the effectiveness such as, equal effectiveness for all, or sufficient effectiveness for the worst-off (ibid.).

The limited source of normativity can also be illustrated by considering the evaluation criteria provided by OECD's Development Assistance Committee (2019) that forms a global benchmark of good evaluation practices for the development programs. It consists in six criteria: relevance, coherence, effectiveness, efficiency, impact, and sustainability. Whereas RFE's are presumably used to study effectiveness only, DAC criteria reflect much broader perspective on evaluation.[21]

So, in this regard, technical norms about means-ends relations are normative, but EBP does not make commitments to specific evaluation criteria other than effectiveness on average. However, this normativity might be jeopardized if the conditions of the technical norms are not being fulfilled (Niiniluoto 1993). Technical norms are always conditional on the current type of situation, in which one founds themselves (ibid.).

This is the point, where the logic of technical norms and normative results of the RFE's often break down as they are criticized for multiple reasons[22]. Cartwright (2007; see also Deaton & Cartwright 2017) has made it clear that extrapolation from one policy circumstance to another require multiple often unwarranted assumptions, such as similar causal structure in the experimental and the target settings. RFE's cannot therefore provide gold standard or shortcut to reveal what works (ibid.). Faverau & Nagatsu (2020a) have also argued that use of RFE's in policy settings require researchers to consider analogous policy scenario in different time, sample, location, in which corresponding intervention will be deployed.

Given researchers' limited knowledge regarding analogous policy scenario or the exact causal structure of the experimental or natural environment, they can be limited to simple descriptive policy evaluation of the historical instance (see Heckman, 2019) instead of more explicit normative guidance. Harrison (2013, p. 108) also notices that researcher can make the inference with as little assumptions as possible and leave the interpretation

[21] However, it has been argued that DAC criteria's one weakness is that it is more suitable for project and programme evaluation than for policy evaluations (OECD DAC 2018, p. 6).
[22] For summary about the relevant literature concerning generalizability issues, see Faverau & Nagatsu 2020a, pp. 14-15.

of what is interesting and valuable to the reader. This seems to suggest that embracing value-neutrality is a viable strategy for randomistas.

In BIE Finland the issue can be illustrated by the differences regarding institutional implementation in the experiment and in possible future policy scenarios. Consider the facts that the tax-model of the experiment was not a feasible feature of the implementable model and the experiment did not have representative sample of the whole population. Therefore, tested treatment did not equal an implementable reform of basic income. This means that it cannot offer straightforward normative guidance of whether it will work in future or not.

The question of how experimental evidence will be useful to design new policy is a more complicated one, however. If one reads carefully the results from the first year of the experiment, one can notice how the focus can shift away from the effectiveness of the treatment to other details. This is one of the defining characteristics of field experiments: one can always try to detect other causal variables influencing the outcome measure (see Haavelmo 1944 as cited in Boumans 2016). According to the research group, it will be more interesting in the analysis to ponder the effects of the monetary incentives and effects of the unconditionality on the outcome variables (VATT 2019, p. 1). In BIE Finland, necessary success factor for this was the possibility to use register data. However, it is not entirely clear, if these results will be useful for planning of the implementable basic income model.

This observation regarding the nature of field experiments and possibility to detect unknown causal variables highlights their adaptability and some similarities with the observational studies, such as natural experiments (Boumans 2016). Among other authors Faverau & Nagatsu (2020a, pp. 18-19) have argued that RFE's can discover "interesting" and "surprising" results and phenomena such as anomalies to current theories. But this feature of field experiment can also introduce other biases and sources of subjectivity. Jacob Stegenga (pp. 23-25) has noticed that if researcher is able to study a range of parameters and choose which one to measure and what kind of comparisons to make, there is a potential for p-hacking. P-hacking can occur in complex data sets and introduce spurious correlations (ibid.).

Given what have discussed above, it seems questionable whether RFE's, in fact, can provide binding and normative results as described by technical norms. On the contrary,

it might be easier for researcher to adopt value-neutrality and leave the interpretation for the policymaker.

As I will argue later, if EBP provides neutral evidence that does not involve normative statements or technical norms, evidence is very hard to interpret by the policymakers and it is likely to undermine the arguments for the practical value of RFE's and introduce another source of subjectivity (5.3.3). If RFE's provide normative evidence, one might wonder if the normative basis of the technical norm is sufficient to provide oughts.

5.2.3 EBP and values: ambiguities

Above, I have discussed the value-free ideal and two specific interpretations of it: impartiality and neutrality. I have argued that even though EBP is not very explicit about it, it seems a reasonable interpretation that it is committed to both impartiality and neutrality. Let me bring up two complications.

The first is that in some instances the distinction between epistemic and non-epistemic values is not that clear even in the arguments of randomistas. This is reflected in the views of Bossuroy & Delavallade (2016) as well as in Banerjee and Duflo (2019), even though differently. It seems a reasonable to ask if their views, in fact, are compatible with impartiality.

Bossuroy & Delavallade (2016) emphasize that RFE's are better policy tools due to their main characteristics, which surprisingly seem to involve both epistemic and non-epistemic values to some degree. They distinguish simplicity, practicality, and relevance that, to a large degree, lean to epistemic virtues of RFE's such as unbiasedness, causality, realism, and lack of theory (p. 149). But they also emphasize simplicity of the results for non-expert audience and easiness of constructing the counterfactual via RFE's (ibid.). This seems to smuggle in some values that are not purely internal for science, but which, perhaps, reflect more the conditions and circumstances of the policy environment.

If my reading of Bossuroy & Delavallade is correct, then impartiality is incompatible with their understanding of EBP. Remember that according to impartiality, research and hypotheses should be appraised by their contribution to epistemic values only, not some non-epistemic values such as feasibility (Reiss & Sprenger 2017). If Bossuroy & Delavallade argue that RFE's are better because they support policymaking in a practical

way that other type of methods and evidence cannot, then the argument is based on other values than ones that are strictly speaking internal for science.

Slightly similar observation regarding the role of non-epistemic values can be made regarding Banerjee & Duflo (2019) as well. They also make room for non-epistemic values to guide the approach of development economists, but very explicitly (2019, p. 8-9). They argue that "it is important that in this project we be guided by an expansive notion of what human beings want and what constitutes the good life" (p. 8). In addition, they urge for "acknowledging the deep human desire for dignity and human contact" (p. 9).

Their approach clearly emphasizes the role of non-epistemic values, but it is less clear how they should be realized when conducting scientific research. They argue that plumbers are guided by intuition, guesswork, and trial and error (ibid., p. 7). This suggests that at least non-epistemic values are, perhaps, important in guiding the more uncertain parts of the approach such as guesswork. But it is unclear whether conducting RFE's remains impartial and value-free endeavour by autonomous and authoritative scientist. It is neither clear would these values concerning the good life influence neutrality or normativity of the results. Banerjee & Duflo (2011, 2019) are clearly driven by the urgency of tackling the global poverty, but it is unclear how these values are realized throughout the research. These observations suggest that clarifications regarding the value-free ideal and the role of non-epistemic values are needed by the proponents of the EBP.

The second complication is that the value-basis of EBP described in the previous section by de Souza & Leao (2019) (libertarian paternalism, leverage, trust in numbers) does not characterize the empirical case study of BIE Finland in all respects.

Firstly, nudging and libertarian paternalism seem to have very little to do with the experimental design in BIE Finland, because it did not involve different service packages designed for nudging. Ex post analysis reveals some differences in the welfare services resulted from the choices of the subjects. However, these inferences could not have been foreseen in the design phase, because statistical strength depends on the size of these groups that resulted in the individual choices. However, the research group have stressed the importance of the research topic for the welfare of the target population (Hämäläinen et al. 2019, p. 12). They were particularly worried about the fact that subjects were very

weakly connected to the labour markets over a preceding two-year period (ibid.). This indicates some type of paternalism and value considerations regarding welfare. However, BIE Finland did not steer the choices of individuals without altering the incentive structure, which is incompatible with the definition of libertarian paternalism (see Grüne-Yanoff, p. 12).

Also, because there was no focus on small interventions or nudges, and because the broader portfolio approach was completely absent, leverage is questionable also in BIE Finland. One the one hand, the idea of scaling-up the solution can be associated at least with the Finnish government's mission of promoting culture of experimentation (Kangas et al. 2019, p. 7). But on the other hand, it is much more debatable whether BIE Finland can leverage basic income policy development. Some commentators have argued that the experiment does not represent major shift towards policy implementation or even significantly re-frame the current political landscape from the existing stable policy paradigm (De Wispelaere, Halmetoja, Pulkka 2018, pp. 17-18). Moreover, as already pointed out, BIE Finland was not a small-scale study, but rather big nation-wide experiment.

Finally, regarding trust in numbers, it should be noticed that the research group did investigate existing experiences and experimental research on basic income, which was considered as an opportunity to guide decision-making (Kela 2016, p. 8). However, the results of the existing literature are downplayed by the observation that there is no directly importable model of basic income for the Finnish society (ibid., p. 8). This highlights different institutional structures, social policy systems, and socio-cultural situations, for instance regarding the labour market, that all have effect for transferring the results from one context to another (ibid.). Moreover, constant feedback loop and learning, perhaps, reflects what research group had in mind when planning BIE Finland (ibid., p. 62), but policymakers' refusal to extend the experiment undermined this intention effectively (De Wispelaere, Halmetoja, Pulkka 2018, p. 15). According to one interpretation, the experiment did not illustrate opening of the new policy window, but rather natural attention shift in politics (ibid., p. 18). However, it showed a commitment to gather and evaluate evidence through an experiment (ibid., p. 17).

This discussion seems to suggest that EBP, in fact, integrates multiple trajectories that can reflect values differently. Not all experiments emphasize nudging, leverage, or

libertarian paternalism to same degree. If I am correct in suggesting that these values can influence the experimental design, it seems plausible to demand clarifications regarding their role in different experimental settings and ask questions about their implications for objectivity.

The discussion in this section casts, nevertheless, doubt on how well the value-free ideal and impartiality can be sustained even in the context of EBP.

5.3 Co-creation and the influence of values

In this section, I will discuss how experiments are being co-created and what role non-epistemic values play in the process. This challenges us to reconsider the value-free ideal and impartiality discussed above.

In her Nobel lecture, Duflo (2019) argued explicitly that their critics often misunderstand how policies are being made. She denies that randomistas would just run the experiment, write the report and leave it there. Rather, policies are being co-created together with governmental institutions and non-governmental organizations (NGO's). Economists' do not work alone or scale up the programmes in isolation. Rather, they are embracing a culture of learning inside governments.

Duflo's views are also well represented in her article "The Economist as Plumber" (2017), in which she argues that economists have both opportunity and responsibility to get all the details of the policies right. As they are involved with the policy implementation they need to be worried even about small adjustments, or" apparently irrelevant details" (p. 4), and complications "some of which may appear to be far below their pay grade" (Duflo 2017, p. 2). Duflo's pragmatic approach highlights the need to reconsider the role of traditional scientist. While she can certainly be credited for reminding about the importance of the programme implementation, constant trial and error, and the need to go beyond mere intuition and assumptions, there is more to be said about the interaction between the researcher and policymaker.

Paul Cairney (2016) has noticed that there is a tendency to discuss the supply-side of the evidence, science, and the demand-side of the evidence, policy, in isolation from each other. According to him, this leads to an inaccurate understanding of the role that evidence can play in the policymaking. While Duflo's metaphor of plumber expands the role of

traditional scientist into the domain of policymaking to cover practical aspects of programme implementation, the supply-side of the evidence and scientific process remain intact and autonomous. Cairney (2016) explicitly argues against two related views of the EBP that he takes to be extremely naïve. On the one hand, he rejects a simplistic account of the EBP, holding that "there can and should be unproblematic link between scientific evidence, policy decisions, and outcomes" (p. 2). On the other hand, he questions the 'policy-based evidence' view that politics is so pathological that no evidence can ever survive the pressures of politicians seeking election, or from the messiness of the political process (p. 2).

In this section, I share Cairney's worries and provide support for both of his arguments from the philosophical point of view.

The EBP can with good reasons be said to assume an authoritative view of science. Despite Duflo might realize the complicated nature of policy influence, she discusses less how scientific evidence is being produced with NGO's or government agencies in various instances. According to de Souza Leao and Eyal (2019, p. 409), EBP represents randomistas as virtuous, although admittedly agnostic, external experts who are looking for right lever of change. As discussed previously in chapter 2, this is explained mostly by distrust towards policymakers. This authoritative view is manifested in the views of Duflo also (2017, p. 15). Duflo stresses that economists make good plumbers by the virtue of their disciplinary training, and modestly states that "this comparative advantage, along with the importance of getting these issues right, makes it a responsibility for our profession to engage with the world on those terms" (ibid.). This is in sharp contrast with the fact that policy scientists are becoming more and more aware about the gap and problematic division of labour between scientists and policymakers (Cairney 2016, p. 54). Most policy theories hold that there is constant interaction between the two via networks (Smith & Joyce 2012, p. 58 as cited in Cairney 2016, p. 59).

Philosophers of social sciences have also examined the conceptual relation between science and policy. Most notably, Heather Douglas (2009, pp. 7-8) has explicitly attacked the isolationist, autonomous and authoritative view of science. One of the main pressing issues is the value-free ideal and whether impartiality is achievable when researchers gather evidence, assess and accept theories (Reiss and Sprenger 2017). Can all this happen without researchers really making contextual value judgments or not?

I fill follow the proponents of the value-neutrality thesis (Reiss and Sprenger 2017), who argue that this cannot happen without involvement of the contextual value judgments. Several philosophers (see Douglas 2009, Khosrowi 2018; Longino 1990,) have argued that non-epistemic considerations, in fact, can and often are integral and constitutive parts of the scientific reasoning, which shows the inadequacy of the value-free ideal. In the following sub-sections I will demonstrate how this plays out in BIE Finland.

Douglas and Longino have recognized the complexity of the distinction between values that are internal to science and the values that are part of the environment (see Douglas 2009, p. 18-19; Longino 1990, p. 4). According to Douglas (2009, p. 90), the main reason to criticize the distinction between epistemic and non-epistemic values is that even in the instances of "good science" non-epistemic values are often smuggled in. For instance, Einstein's theological views influenced his epistemic choices in physics (ibid.). Douglas (2009, pp. 18-19) also argues that the distinction between constitutive and contextual values makes it harder to recognize the boundary between science and society. Longino's (1990, p. 4) discussion about constitutive and contextual values started with the question "can the distinction, as commonly perceived, be maintained". Her point was to highlight how constitutive and contextual values are in dynamic interaction and that scientific practice requires such interaction (Longino 1990, pp. 4-6). Difficulties with the value-free ideal inspired both Longino and Douglas to search for alternative and refined concepts for objectivity (Reiss & Sprenger 2017).

While the arguments against impartiality are many (see Reiss & Sprenger 2017), *inductive risk* is the main reason why the value-free ideal has been abandoned (Douglas 2016). Inductive risk is a risk of being mistaken when making decisions in science (Douglas 2016). It consists in legitimate disagreements over what counts as sufficient evidence in science and implies the need of judgment and values throughout the scientific process (Douglas 2016). The earliest arguments have been found in the writings of Rudner (1953), Hempel (1965), and William James (1896), and inductive risk has been much discussed topic in the 21st century (see Douglas 2000, 2009, 2016; Elliott and Richards 2016; Risjord 2014).

One of the early versions of the basic argument is that researchers necessarily make value judgments when testing hypotheses and choosing theory (Rudner 1953, as cited in Risjord 2014, pp. 20-22). Rudner pointed out that hypotheses can only be proven more or less

probable rather than definitely as true or untrue. This requires researchers to consider the level of accepted uncertainty, to select the proper p-value, and the level and magnitude of the risk that is accepted. Researches must assess the acceptability of two different types of error that cannot be reduced simultaneously because they are often inversely related: false positives (in statistics: type I error) and false negatives (in statistics: type II error). So, here is always a possibility that results are either over- or underreported. Because researcher must make a choice between these different types of errors, she is forced to consider the costs of mistake. This opens a door for non-epistemic value judgments to influence the decision, which shows that they are constitutive to accepting and rejecting hypotheses.

Heather Douglas (2000, pp. 563-565) has argued these value choices do not occur only when theories are accepted or rejected, but throughout the internal stages of science. These value choices occur in the methodological choices, in gathering and characterizing of the data, and in interpreting the data. The methodological decisions concerning the design of an experiment, or the statistical method for analysing and processing the data, cannot be done without considering the consequences of error (Douglas 2000).

Consider the following observation by Roth (1988, p. 1023 in Santos, p. 93): "there is room for an experimenter's prior beliefs about the likely outcome of the experiment to influence the outcome, through these design decisions". According to Roth, scientists must make decisions about the level of many parameters in the experimental setting.

Douglas (2000, pp. 565-578) also illustrates her argument with the example of the dioxin studies and shows how selection of the statistical significance level, characterization of the data, background assumptions, and model choices can lead to under- or overregulation of the chemicals, having varying costs for public health or regulated industries. According to Douglas, the fact that the choices of the researcher yield non-epistemic consequences, require non-epistemic values to be present in the decision-making. Non-epistemic values are therefore necessary part of the science.

There are two important and interrelated aspects in inductive risk that scientist must consider: the magnitude of the risk and valuation of consequences (Douglas 2000, p. 565). Scientist has a moral responsibility to consider the consequences of her decisions but also potential of the epistemic risk (Douglas 2009, pp. 73-75). The proper understanding

concerning the role of values in science should have therefore foundations in "moral responsibility" and "proper reasoning" (Douglas 2009, p. 20).

This highlights how epistemic considerations of error and valuation of non-epistemic consequences are profoundly tied together in inductive risk. However, it is not clear what consequences inductive risk bears on objectivity in different circumstances. Douglas has noted on several occasions that inductive risk has no yet clearly understood implications for objectivity (Douglas 2000, p. 578) and involvement of ethical and social values "need not threaten the integrity of science in policy" (Douglas 2009, p. 154). Although it raises the difficult question concerning how to construe objectivity of science among the other complex issues regarding authority and accountability of science (Douglas 2016). As she states: "we cannot expect *perfect* foresight and prediction. But we should expect *reasonable* foresight and care from our scientists" (Douglas 2009, p. 67, original italics).

Several authors and case-studies have examined inductive risk in various fields such as pharmaceutics, physics, and psychology, and established its connection to low evidential standards of the scientific community (see Elliott and Richards 2016, pp. 5-6; Stegenga 2016). This highlights how inductive risk is not only a matter of ethics, but also an epistemic issue dealing with our expectations of sound science. Some of the case studies deal explicitly with the demand of prioritizing RFE's (see Stegenga 2016; Bluhm 2016). Stegenga (2016), for instance, shows that there are many other important matters, in addition to the number of RFE's, that should matter when making decisions about drug regulation based on pharmaceutical evidence, and he criticizes epistemic standard of US Food and Drug administration for the lack of "any reasonable norm of evaluation" (p. 23). Bluhm (2016, pp. 10, 207-208) also concludes that examining inductive risk in clinical trials reveals that there should be more focus in assessing the type of evidence, rather than how much evidence there is.

Inspired by the case studies examining inductive risk[23], and especially Douglas (2000), I will discuss three routes how ethical and political values can become constitutive values of the experimental research in policymaking as illustrated in BIE Finland. These are: (1) decisions of the experimental design, (2) selection of a theoretical content, and (3) interpretation of the results.

[23] See book *Exploring inductive risk: case studies of values in science* (2016) by Elliott & Richards (eds.).

As I will argue, these routes exemplify points of decisions when scientist must assess the magnitude of risk as well as non-epistemic consequences of their choices. Therefore, they present epistemic risks of varying degree to the EBP, potentially increasing subjectivity of the research. In this thesis, I will focus mainly on their epistemic aspect and implications for the evidential standards and discuss less non-epistemic consequences from the moral perspective. I will hence illustrate underlying epistemic risks and opportunities for the research, pondering their implications for objectivity and sound science.

5.3.1 Experimental design

Case studies in various fields have documented a wide range of technical decisions that have been associated with inductive risk. These include, for instance, the level of statistical significance (Douglas 2000, pp. 565-569; Andreasen & Doty 2016), characterization of the data (Douglas 2000, pp. 569-572), background assumptions and models (Douglas 2000, 573-577), impact and outcome measurement (Andreasen & Doty 2016, Stanev 2016), operationalization (Andreasen and Doty 2016), trial quality (Plutynsky 2016), and baseline mortality (Plutynsky 2016).

In this section, I will discuss some technical decisions of the experimental design in BIE Finland, illustrating inductive risk of the experimental research in policymaking. They involve decision over (i) the selected model for testing, (ii) the sample size, (iii) the level of primary treatment variable, (iv) the subject population in the experimental setting, (v) the level of variation of the treatment variable, and (vi) the selected treatment variables. These are points of decisions, in which researcher face uncertainty and the risk of error, because of co-creation of RFE's. As I will illustrate, they also open routes for non-epistemic values to influence these decisions, whether legitimately or not. Hence, they require a careful judgment of the researcher and consideration of epistemic and non-epistemic consequences of these choices. I will also argue that they demonstrate points of decision, in which evidential standards could potentially be clarified and improved by the EBP.

Let me start with the selection of the basic income model (i), which was based on the comparison of alternative basic income models and their cost-effectiveness calculations. Many authors in the economics literature have criticized high costs of conducting RFE's

and noticed that one should run cost-effectiveness calculations prior to experiment to justify the costs of experimenting or different policy alternatives (see Deaton & Cartwright 2017; Dolan & Galizzi 2014, pp. 728-729 Harrison 2014).

The research group in BIE Finland seemed to follow this recommendation when they were selecting the basic income model for the experimental testing. The basic income model was justified, at least partially, by some type of cost-effectiveness calculus.[24] Calculation compared different models of basic income and their pros and cons, based on both epistemic and non-epistemic values.

As already mentioned previously, the suitability of the five examined models (full basic income, partial basic income, negative income tax, participation income, and universal credit) for the experiment were compared in the pre-study based on the broad comparison of pros and cons in six dimensions: (1) economic incentives for work, (2) bureaucracy and administration, (3) poverty and income gaps, and participation, (4) feasibility, (5) support base, and (6) feasibility for testing (Kela 2016, pp. 53-57). Dimensions (4) and (5), feasibility and support base, concern practical aspects of policy environment and represent non-epistemic values underpinning it. Dimensions (1) - (3) concern the aims of the experiment: from providing economic incentives, to reducing bureaucracy, and tackling poverty. These are also non-epistemic values, guiding the selection of the model.

Assessment of the sixth dimension, feasibility for testing, seems to represent more explicit epistemic values and issues that the research group associated with the specific models. Even though it was not a very thorough analysis, it involved research group's comments on credibility, reliability, extrapolation, as well as practical requirements for data collection. All reflecting epistemic values.

As all comparisons along these six dimensions, feasibility for testing also, compiled together more or less different observations regarding the different models. Regarding feasibility for testing, all observations regarding different models were different in nature, dealing, for instance, with the sample size (full basic income), required income register for real-time income monitoring (negative income tax), local level and difficulty of extrapolation (participation income), and the outcome measure (universal credit). All except one being cons, and one of the models lacking any pros and cons (see image 1

[24] I will discuss later the argument that selection of the model was not, in fact, affected by the cost-effectiveness calculation of different models.

below). Although elsewhere, partial basic income model was appraised mainly by the bigger sample size and easier generalization (Kela 2016, p. 54).

	Full basic income (1,000+ euros)	Partial basic income (<800 euros)	Negative income tax	Participation income	Universal Credit
Feasibility for testing	(-) Would make the experiment less credible scientifically as the sample size would, for reasons of cost, remain smaller than for lower basic income levels.		(-) Would require a national income register for real-time income monitoring and monthly payments of guaranteed income based on personal income reporting would weaken the scientific credibility of the experiment.	(+) Could be tested in small scale and at the local level. (-) Reliability and extrapolation of the results is difficult. (-) Would not allow testing of the effects of unconditional basic income. Because of its conditional nature, would be better suited as a model for a research group developing participatory social security. (-) Constructing a payment platform for the experiment would be difficult. (-) Organisation would require additional resources. Who would be responsible for the organisation?	(-) Would require an income register; testing would not be currently possible. (-) Because of conditionality, would not allow the testing of the effects of unconditional basic income.

Image 1. Feasibility of different basic income models for testing (Kela 2016, p. 57).

Cost-effectiveness calculus demonstrates a deliberate attempt to clarify both epistemic and non-epistemic consequences of the model choice. It shows how epistemic and non-epistemic values both can influence selection of a model and how the scientific approach in BIE Finland was not impartial and value-free. Non-epistemic values of reducing bureaucracy and poverty were deeply entangled with epistemic values as characterized in feasibility for testing, which both were constitutive for the whole approach.

I take this to be a legitimate instance of non-epistemic values influencing scientific decisions, not threatening objectivity, because of its close connection with selection of the research problem. Whether these six dimensions represented appropriate and relevant characterization of the five models is a more difficult question that requires elaboration of the political and ethical theories and is beyond this thesis. It nevertheless illustrates how non-epistemic values are needed when making decisions over the experimental design in policymaking.

Notice also that it is not self-evident how much weight to any given criterion should be given and why. Even if one model were the best option along other five dimensions, if it is completely unfeasible, it might not qualify for the experiment, and legitimately so. Therefore, the comparison should be understood, even ideally, as somewhat balanced or neutral assessment of pros and cons, leaving the final decision and more explicit valuation of the dimensions to the decision-maker. This seems to imply the division of labour

between researcher and policymaker regarding describing the options and making decisions.[25]

However, as one of the dimensions, feasibility for testing, involved epistemic considerations regarding feasibility for testing, weighting these pros and cons should probably not be the sole responsibility of the policymaker but require active participation of the researcher as well. If support base or economic incentives were weighted more in decision-making, it could mean less weight on feasibility for testing (epistemic values). It should be added that the research group did make their own recommendation for the most suitable model as well, but their recommendations were not followed (see Kela 2016, pp. 5-6, 58-62). Nevertheless, cost-effectiveness calculus provides an interesting framework for the co-creation of the experiment, which challenges us to consider the roles and responsibilities of researcher and policymaker as well as epistemic and non-epistemic values in the process.

Let me add one more complication to the selection of a model. As the assessment of the different models is guided by the extensive background knowledge (Kela 2016, pp. 55-57), microsimulations (see Kela 2016, p. 16-49), and different types of evidence, it presented another avenue for the risk of error, opening another route for non-epistemic values to influence the characterization of the models via inductive risk. While models were compared by their respective features regarding the above-mentioned dimensions, in some instances, the comparison also involves expectations regarding the models' capability to achieve these aims effectively. For instance, the research group states that full basic income model "would be effective tool for reducing poverty", some models "would benefit civil society organisations" and partial basic income model "would weaken the livelihood of individuals eligible for social security supplements" (ibid., p. 56).

This assessment of pros and cons, therefore, require an assessment of different types of evidence available for the research group, not only consideration of the features of the different models. This makes cost-effectiveness calculation itself a subject for inductive risk. Researchers must choose the level of required evidence for all their observations, which opens the door for non-epistemic values to influence characterization. However,

[25] See also Douglas' (2009) example of division of labor in risk management to descriptive *risk analysis* and decisions over accepted level of risk, *risk management*.

as already argued, it is not clear what implications these risks of error would have for objectivity of the research or responsibilities of the scientists.

When reading the cost-effectiveness calculus, it becomes clear that this kind of comparison is extremely challenging to make, especially given the limited time for planning. The researchers are forced to make decisions about their approach based on their existing knowledge in respective fields. Even though this reasoning is not fully transparent in cost-effectiveness comparison, I do not consider this type of inductive risk as realized in the selection of a model as a threat for objectivity. However, it demonstrates that scientific decisions in EBP are not based on the epistemic values only and non-epistemic values can influence the research on multiple levels.

However, there is a more direct threat for objectivity with non-epistemic values that one deriving from selection of a model. One could argue that non-epistemic values directly interfered with the selection of the model, bypassing the whole cost-effectiveness comparison, as there was no real choice over the most appropriate model. Despite comparing alternative models for basic income, researchers could not do all the technical choices over the experimental design on their own in BIE Finland. Many design choices were driven by budgetary, legal, institutional, and political reasons (De Wispelaere, Halmetoja, Pulkka 2018, p. 16). Even according to the value-free ideal, this would count as direct violation of impartiality, if the model choice was understood as choice concerning internal stage of the science, data gathering, and not as selection of a problem.

So, despite cost-effectiveness calculation discussed previously, all recommendations of the research group were not implemented for various reasons (see Kela 2016, pp. 58-62). In the published postscript (ibid.), the research group explained why their recommendations were not adopted and responded to the criticism towards the experiment. One of the main reasons (and complaints) concerned the financial side of the experiment (see Kangas 2016; Kela 2016). The allocated budget for the experiment, 20 billion euros including administrative costs, was relatively small and did not meet the research group's expectations regarding the size of the experiment (Kela 2016, pp. 13-14).

The research group also faced another closely related constraint that was presumably intended to specify the research problem, but which had also implications for the experimental setting. The proposed and tested model of the basic income had to be cost-

neutral. This meant that the proposed and tested model for basic income could not have increased the total government spending on social welfare assistance. In other words, the research group had to figure out how the proposed basic income model would be financed via taxation (see Valtioneuvoston kanslia 2016, pp. 188-189). Partial basic income at the level of 550 euros, for instance, required collection of 11 billion euros more (ibid. p. 73). Regarding households, cost-neutrality also meant that the net income of the people would not change significantly in the experiment (ibid., pp. 73, 208-209).

Allocated budget for the experiment and the demand of cost-neutrality are obviously a financial (non-epistemic) values, influencing selection of a research problem, but they also had an impact on the gathering of evidence. They had a direct effect on the size of the sample size (ii), the level of treatment variable (iii), and the selection of the subject population (iv).

Firstly, as the head of the research group, Olli Kangas (2016), notices the sample size had to be reduced from initial 10 000 to 2000 persons. Secondly, as the level of basic income paid for the target group was also limited by the overall budget, the level of the treatment variable was also affected by the budget. Senior Specialist at the Prime Minister's Office, Markus Kanerva (as cited in Jacobin 12.1.2019), commented for instance that

> "It was relatively easy actually to choose a partial basic income. Full basic income seemed unfeasible because it would be very costly, and we were given the twenty million by the [government] for two years together, so our sample size would be pretty limited with a high-income amount. Plus, there was also the demand that the model that would be tested would be somehow budget-neutral. So, in that sense, it was lowered to a partial basic income […]".

Thirdly, the budget constraint influenced the selection of a subject population and lead to the exclusion of the more expensive target groups, such as students (Kela 2016, pp. 60-61). Although the initial idea of the experiment was to test basic income with the entire non-retired adult population (Kangas 2016), the smaller sample was justified by its cost-effectiveness, ease of implementation in the given institutional setting and in drafting the legislation, by speedy and efficient sampling, and tight timeline (Kela 2016, pp. 60-61).

The demand of cost-neutrality also had complicated consequences for the whole research problem, because cost-neutrality implied that it was very hard to increase the economic incentives for work and decrease participation tax rate in the experiment (Ville-Veikko

Pulkka as cited in *Politiikasta* 3.3.2018). Microsimulations showed that in the selected model of basic income the marginal tax rates and participation tax rates would have been lower than in the current system (Kangas 2016). This together with the withdrawal of the tax administration lead to definition of basic income as tax-exempt in the experimental setting, decreasing its cost-neutrality, and creating a mismatch with implementable future basic income model.

Let me add two more constraints representing non-epistemic values and which complicated the design decisions and raised the risks of error in the experimental setting. One deals with the level of variation in the treatment variable and the Finnish constitution (v) and the other one with the selected treatment variables and the Finnish Tax Administration (vi).

Firstly, the Finnish constitution, together with the research group's preference of avoiding voluntary experiments, effectively ruled out the possibility to test multiple levels of basic income (Kangas 2016; Kela 2016, p. 60). This made it impossible to vary the level of treatment variable in the experimental setting. It also fixed the level of basic income to 560 euros, which equals to the labour market subsidy and the basic unemployment allowance in the current system. The Finnish constitution's worry regarding treating people differently thus affected the design and possible inference constraining the manipulation of the size of the financial incentive.

Secondly, the Finnish Tax Administration did not participate in drafting the bill, despite the initial idea of examining the level of basic income together with different tax models and tax rates (Kangas, 2016). Some authors said that "the tax authorities were not interested a bit in the experiment" (Markus Kanerva as cited in Jacobin 12.1.2019). This meant that the experiment would occur within the existing tax system, and tax models and rates could not have been varied in the experiment either (Kela 2016, p. 60). This constrained further the research group's options for achieving cost-neutral model as they had suggested several cost-neutral models, which turned out have a huge administrative cost and impossible timeline for implementation and legislative work (ibid., p. 59).

Regarding the Finnish Tax Administration, there has also been a speculation about the role of political resistance and opportunism by the responsible Minister of Finance, Petteri Orpo. Allegedly, Minister Orpo "used his powerful position to prevent Kangas and his

design team from incorporating any tax component into the experiment" (Jacobin 12.1.2019).

Regardless of the role of hidden political agendas or sabotage, this shows how non-epistemic values and the lack of partnerships constrained the space for the experimental design and made it impossible to vary either of the potential treatment variables: the tax rate or the level of basic income.

To summarize the discussion so far, all this shows how epistemic and non-epistemic values both are constitutive for the experimental design in policymaking. Secondly, it demonstrates specific issues that require a careful judgment from the researchers. Sample size, subject population, model choice, and treatment variables (selection of treatment variables, the level of treatment variable, and the variation of the level of treatment variable) are all extremely important aspects of the scientific inference, having potential consequences for reliability and objectivity of the study. Researchers are therefore forced to consider the risk or error, but also non-epistemic consequences of their choices for instance, for policymaking.

But, what consequences all this bear on objectivity of the BIE Finland? After the initial enthusiasm of the experiment, the perception has been much more modest in the public and, in several instances, highly critical. Several articles in the media (see Jacobin 12.1.2019; Kangas 2016; Politiikasta 3.3.2018; The New York Times 20.7.2017;) have declared it a failure, pointing towards bureaucratic obstacles of timeline and budget, partisan politics and resistance among civil servants, as well as politically defined research problem, just to name a few reasons for criticism. The postscript published by the research group also mentions some of these criticisms: unsuitable target population, outrageously expensive or meagre experiment, small and skewed sample, and unsustainable model (Kela 2016, pp. 59-60).

The head of the research group himself, Olli Kangas, also demonstrated frustration with the design of the experiment in his public lecture[26] "How to plan a successful field experiment – and how to destroy it". He was especially frustrated with the planning process and some unfavourable decisions by policymakers, such as the budget constraint limiting the sample size of the experiment. Despite the provocative title, his conclusions

[26] 13.8.2018, Public lecture, The University of Helsinki, Swedish School of Social Sciences, Festhall

regarding the reliability of the experiment were more moderate, while he admitted that the results could be somewhat useful in future in the hands of the academic community. Also, elsewhere[27] in public he has been defending the reliability of the results.

So, what should we think of all of this? Were these matters legitimate research choices or did political values unduly interfere with science and its methodological choices?

One could argue that even though financial values and demand for cost-neutral model primarily influenced selection of the research problem, they also had indirect consequences for the strength and outcomes of the study at least, as far as extrapolation for the population level is being considered. This was due to limiting the treatment group to a specific sub-population. Moreover, the interpretation of the Finnish constitution ruled out the possibility of varying different levels of treatment, which also had consequences for the strength of the study regarding generalizability of the results concerning incentive effects.

But notice that even these observations would not necessarily undermine objectivity of the research, even though there are potential epistemic risks for extrapolation or generalization. As discussed in the previous chapters, however, it is not clear to what degree extrapolation and generalization are required from RFE's, if RFEs are understood as tools for simple policy evaluation of any historical intervention, as means to estimate population level average treatment effect given the characteristics of the sample, as offering interesting or surprising results, even though only explorative evidence, about bureaucracy traps? This requires EBP to consider the aims and values more explicitly and to step outside of the narrow mechanical objectivity of RFE's.

5.3.2 Theoretical content

In this section, I will discuss the second route for non-epistemic values to become constitutive values of the research as illustrated by BIE Finland. This is through the theoretical content (see Risjord 2014, p. 20). By theoretical content Risjord (ibid.) refers to the products of the research and the related discussion concerning value neutrality and normativity of the results. As he puts it "value neutrality is the thesis that social scientific theories should describe facts, not make policy recommendations" (ibid., p. 24).

[27] Twitter

This seems like a wrong dichotomy as the EBP is entangled with both. As I already discussed in section 5.2, I think it is somewhat ambiguous whether EBP provides normative or descriptive results. However, what is here more at stake than the normative character of the scientific product, is the theoretical ingredients that are used to achieve these products, whether they involve oughts or not. As was discussed in the previous chapters, EBP is extremely keen to avoid theoretical assumptions or grand theoretical frameworks concerning what works. But, as I will show next, theoretical content, if understood correctly, is an important and relevant matter to experimental research. It opens another route for non-epistemic values to become constitutive values of the research, which, in my view, is not only a threat, but an opportunity to increase objectivity.

Some economists (Dolan & Galizzi 2014, pp. 729-730; Harrison 2013) have become increasingly aware of the need for specific theoretical content, such as welfare impacts, when designing experiments and normative policies. Harrison distinguishes several types of inference that experimenters can make and argues that many of them, in fact, should require specific theoretical content (Harrison 2013, pp. 105-111). This theoretical content seems to involve both epistemic and ethical values to some degree and deal with certain type of outcome measures.

Firstly, experimentalists can study and evaluate the welfare effects of a treatment in order to do a cost-benefit analysis of policies (Harrison 2013, pp. 105-107). Harrison argues that in order to favour any one policy alternative over the others, one must understand the welfare impact of the policies (ibid.). The welfare impacts make the results and comparison of policy alternatives meaningful. In BIE Finland, welfare impacts were incorporated into the design of the Finnish Basic Income Experiment partially. Despite the primary aim of the experiment was to examine the effect of the basic income on the labour market supply, welfare impacts were studied in the target and control group. Welfare impact was not, therefore, deeply integrated into the experimental design in the form of specific treatments or designed service packages, but it provided an additional source of information regarding the effects of the treatment, gathered via surveys.

However, according to Hiilamo (8.2.2019), survey results in BIE Finland were not reliable for three reasons. Firstly, no baseline survey was integrated into the research, which means there is no way to know if the results were due to basic income. Secondly,

subjects' awareness of the public debates, concerning for instance simultaneously introduced activation policy, can have biased self-assessment of subjective wellbeing. Thirdly, low response rates in the treatment group (31%) and in the control group (20%) decrease reliability.

Despite these epistemic disadvantages, I will argue that examining the welfare impact presented an opportunity to increase validity and relevance of the evidence provided by BIE Finland, by bringing in additional, value-laden, source of information that can potentially reduce the risk of error.

Notice that Hiilamo (ibid.) did not reject the usefulness of the survey results completely, but claimed that "Despite shortcomings, the survey results are interesting enough to keep up interest in basic income, especially among those who see basic income more as a social justice issue than an instrument for activating the unemployed people."

This nevertheless demonstrates how moral and ethical values can play a legitimate role as a theoretical lens and become constitutive values of the research. Additional sources of information can often reduce inductive risk regarding confirmation and rejection of the hypotheses that the researcher faces and increase objectivity of the results. However, in BIE Finland, this opportunity was not fully seized, as the survey suffered from serious epistemic disadvantages. According to Dolan & Galizzi (2014, p. 729), integrating welfare into RFE's could also potentially help understand further theoretical questions concerning risk, time, and social preferences. Whether this will happen in BIE Finland, is yet to be seen.

The second inference that Harrison (2013, pp. 109-110) discusses is that experimentalist can study the distributional impacts and the winners and the losers of a certain policy. He argues that if we care about the distributional impacts, we want to identify the individuals that are losing as a consequence of the policy and simulate the policies that could potentially mitigate losses.

Donal Khoswrowi (2018, pp. 9-20) has argued that if RFE's lack information on group heterogeneity, they are not suitable to examine distributive consequences, but usually report effectiveness on average. This, according to Khosrowi, illustrates a trade-off relation between RFE's epistemic values underpinning the methodology and non-epistemic values related to equality and distributional effects. Khoswrowi discusses sub-group analyses that can be used to provide information about heterogeneity, but notices

that their results are usually explorative due to limited sample and they can introduce spurious correlations and questionable identification assumptions. Therefore, they do not enjoy the same epistemic authority than RFE's, but are lacking in unbiasedness, causal conclusions and methodological rigour.

In BIE Finland this type of theoretical content, concerning distributional effects, were partially analysed via subgroup analyses. Various distributional effect could be analysed with the help of the extensive register data, but they suffered from the same observations made by Khosrowi: the sample size for the proper subgroup analyses were too small and imprecise (VATT 2019, p. 18). Thus, there were most significantly epistemic opportunities in this regard, but also risks because the opportunity was not again fully seized. Nevertheless, subgroup analyses in BIE Finland could provide explorative evidence that can, for instance, guide future research.

Given these two examples of welfare impacts and distributional impacts, one could argue that when designing normative policies, we might want to include specific theoretical packages in the experimental design. To me, it seems clear that welfare impacts and distributional impacts bring in non-epistemic values that can be constitutive for the experimental research. They are theoretical lenses that researcher might want to apply when studying certain treatments or phenomena.

These theoretical lenses are also related to the endemic risk of error in scientific endeavour. In studying a subject with such long history in Finnish politics as basic income, researcher and policymaker face a difficult choice of selecting outcome variables under investigation. As already discussed, the primary aim of the experiment dealt with the labour market effects. But selecting theoretical approach necessarily narrows down the subject, and hence the researcher must consider the cost of mistakenly selecting a bad theoretical orientation. However, this choice is clearly related with the research question and selection of a problem. Applying these additional theoretical lenses in the research can help the researcher to decrease the risk of error by providing additional evidence on welfare impact or distributional effects. But this decision to study these issues is likely to be influenced by non-epistemic values concerning these theoretical packages, basic income, and the ideals concerning scientific research. Regarding these theoretical lenses, there are, therefore, both epistemic risks but more significantly opportunities that require a careful attention from the researcher.

However, it is again hard to argue that specific theoretical choice of studying the welfare, distributional impacts, or the lack of them would directly violate objectivity of the research. Nevertheless, it shows how non-epistemic values, indeed, play a role in making these methodological decisions and can become constitutive of the research, improving the research setting. When studying major societal reform like basic income, it seems plausible to argue that if we take inductive risk seriously, non-epistemic consequences of these choices should also be considered when making these methodological choices.

It should be noticed that not adopting any specific theoretical lens described above, does not mean that one could completely ignore this matter on non-epistemic values and that they would not play any role whatsoever. Harrison (2013, p. 105) has discussed the possibility of conducting experimental research without explicit theory. In this instance, experimentalist can estimate the average effects of a policy on directly observable outcomes. This type of inference is the only one that does not require explicit theory. But he criticizes this approach and its reluctance to make any structural assumptions about why something works, and the limitations to the average effects (Harrison 2014, 754).

This shows how the lack of theoretical content can be criticized for epistemic reasons, because "we need to know more than just the average effect" (Harrison 2013, p. 110). These arguments appeal to epistemic values and present reasons to apply various theoretical lenses. According to Harrison, researcher must consider causal mechanism that might interact with the treatment in heterogeneous settings (2013, pp. 110-111). Researcher neither needs to be limited to the evaluations of directly observables, because there are important and interesting questions for scientists regarding individual and social welfare, which all involve latent constructs (ibid.).

Examining welfare and distributional impacts, and causal mechanisms, are all epistemic matters, but they are in complex interaction with non-epistemic values, especially when making policies.

5.3.3 Interpretation of the results

In this section, I will discuss two ways how BIE Finland can be argued to lead speculative ex post analysis allowing non-epistemic values to influence the interpretation of the results. First concerns (i) the results from the RFE's, and second concerns (ii) the

comparison of two different types of evidence that can complicate confirmation and rejection of the hypotheses. Let me start with the first one.

Harrison (2014, p. 755; 2013, pp. 107-109) has argued that in many instances of RFE's, researchers are left with open ending *ex post* analysis on the reasons why certain behaviour occur, leading to "casual theorizing**"** and "even more casual behaviourism". Which, "is like doing brain surgery with a divining rod" (p. 108). He argues that in order to design normative policies, one needs complementary understanding about why certain behaviour arises, which is not usually supplemented.

In BIE Finland, such complementary understanding regarding mechanisms and reasons of certain behaviours is largely absent. Risk attitudes, time preferences, or subjective beliefs were not integrated into the experimental design. It is not yet clear, how survey results will be useful in the final analysis.

Although it is not known yet what the ex post analysis will look like, results from the first year of the experiment illustrate that Harrison's worry about speculative analysis might be relevant. For instance, the research group reports the observation that the subjects applied eagerly for the unemployment benefits, which was due to additional monetary incentive[28] (Hämäläinen et al. 2019). According to the research group, this indicates that "active labor policy was not experienced so disgusting that people had chosen to get rid of it" (ibid., p. 31). The research group also remarks that "voluntary participation might be experienced differently than mandatory" (ibid.). Harrison (2013, pp. 107-109) directly attacks on this kind of speculation about people's subjective beliefs and experiences, because of lacking data, and as he argues, it is a tautology to infer from the choices any preferences, because in economics preferences are defined by people's choices. It is, therefore, troublesome to give labels such as 'disgusting', without providing an additional proof. However, further analysis with the help of the register data will be hopefully published after the experiment and survey data can potentially help to fulfil the gaps and help us understand why certain behaviour occurs.

Let me now discuss the second concern regarding the interpretation of the results, which derives from two types of available evidence; and relates with acceptance and rejection of the hypotheses (ii).

[28] provided by the child increase of the unemployment benefit (Hämäläinen et al., p. 31).

In the Finnish Basic Income Experiment, survey data aimed at supplementing the results with analysis of the welfare impact, in contrast to the narrower labour market effects, studied via the actual RFE and average treatment effect (ATE). However, survey results on the welfare impact and the results from the actual RFE on the labour market effects did not align for the same conclusions (see Hämäläinen et al. 2019; Kangas et al. 2019). Whereas survey results reported for the subjects' improved welfare, RFE did not report any significant change in the labour market situation (ibid.). These two pieces of evidence concern different effects of the basic income and are not in direct contradiction with each other, but they also seem to suggest different conclusions regarding the usefulness of the basic income in the Finnish society. The comparison of these types of evidence, thus directly evokes a considered judgment regarding the quality and relevance of all available evidence, as described by Reiss (2014), and explicit value judgment about desirability of the respective effects. One could argue that such value judgment is not researchers to make as they only present their findings and let policymakers to make these valuations. Thus, researchers are neither forced to compress acceptance of these two different findings into one, because the research consisted in two hypotheses, not one. While I think this is true, I think there is also certain ambiguity regarding the promises of the EBP to dictate 'what works', which suggests that this kind of comparison could, perhaps, be made. In section 5.2.2, I discussed value neutrality and suggested that in fact EBP is much better equipped to stick with descriptive results and not provide any normative recommendations. However, given the ambiguity of the EBP and the lack of explicit discussion about the value neutrality, I have chosen to discuss the issue here as a reminder. Also, because of the research is being conducted outside the academia, where might be misconceptions and different expectations regarding the recommendations that researchers will deliver after the study, the issue is not without relevance.

Interpretation of the results and making conclusions about 'what works' presents, thus another case of inductive risk. Emphasizing survey results on the welfare impact could push evidence towards false positives (false positive effect of the basic income for the subject's welfare) whereas emphasizing the lack of significant labour market effect could push it towards false negatives (false negative effect of the basic income for the labour market supply). This allows non-epistemic values, and considerations regarding non-epistemic consequences, to influence the interpretation of the results; and potentially acceptance and rejection of the hypothesis.

This is further complicated by the fact that both types of evidence can be argued to be compromised and to be insufficient to some degree. Recall the difficulties discussed in section 5.3.1, dealing with sample size and subject population, and ultimately with generalizability and extrapolation. However, surveys also suffered from epistemic drawbacks as was discussed in section 5.3.2. Moreover, subgroup analyses were also carried out to study distributional effects, but they do not fulfil the evidential standards of the EBP as they can provide only explorative evidence.

Concluding analysis will, thus require a careful judgment concerning relevance and quality of different types of evidence**.** This involves judgments concerning epistemic values but will also demonstrate how research team will ultimately be concerned with uncertainty and risks of error, present in the study setting. This opens the door for non-epistemic values to influence conclusions and possible recommendations. However, epistemic values and related methodological concerns about the quality of the evidence are deeply entangled with non-epistemic values related to (i) reported welfare impact, (ii) reported labor market impact, and to (iii) respective non-epistemic consequences associated with over- and underreporting of the respective effects of the basic income. It is therefore also debatable whether policymakers could make respective value consideration (see Douglas 2009, pp. 44-86, for the argument concerning the moral responsibilities of the researchers).

My observation, therefore, resembles one made by Khosrowi (2018) that RFE's need to be supplemented with other methods, such as subgroup analyses, or survey data as in BIE Finland, which do not satisfy the evidential standards of EBP. Therefore, when making conclusions of what works, researchers are forced to balance between different epistemic and non-epistemic values. This observation also raises a worry noticed by Khosrowi (2018) regarding how other pieces of evidence can be disregarded or neglected when conducting RFE's. But, as I argue, this simply depends on research teams' trained judgment and epistemic and non-epistemic values.

In both instances of the interpretation of the results, there are epistemic risks present. First instance illustrates insufficient knowledge about the mechanisms behind the occurring behavior and subjective speculation by the researcher (i). Second instance illustrates the challenging task of assessing the different pieces of evidence together (ii), which requires consideration of their relevance and the quality. As all confirmation and rejection of

hypotheses it carries an epistemic risk of making the wrong conclusion or recommendation, which in turn can lead to harmful non-epistemic consequences.

6. CONCLUSIONS

I have so far discussed EBP's preference to conduct randomized field experiments in policymaking and investigated their claim that RFE's provide better evidence than other available methods in terms of their objectivity.

I have argued that we cannot unequivocally rely on the evidence from RFE's in policymaking, because of its narrow emphasis on mechanical objectivity. Despite of EBP's distrust towards their fellow economists' theories and methods (2.2.1), and the lack of evaluation of governments' policies (2.2.2), the solution is not to uncritically rely on RFE's. They do not provide great tools that would serve equally policymakers and researchers (3.1) or simply reveal brute facts and how the world lies (3.2). Rather they rely on strict mechanical protocols and methodological norms (3.3.) that are largely insufficient to ensure that the evidence from RFE would effectively avert our beliefs.

As my discussion about BIE Finland in chapter 4 illustrates, the features and epistemic values of the approach can be compatible with multiple epistemic aims, to begin with. Comparison of the features and biases of two waves of RFE's also show that there is room for value choices regarding what counts as a successful RFE. Whether RFE's function as simple testing procedures, policy evaluations of the historical or implementable policy, or an attempt to extrapolate, affect their epistemic authority.

I have also argued that respective guidelines of EBP, do not consider the fact that the quality and nature of the evidence can change in the process due to the co-creation of the evidence-based policy. Policy environment and co-creation introduces new epistemic risks, opportunities, and complex value trade-offs between epistemic and non-epistemic values. What appears to be a contextual value affecting the selection of a problem, can have far-reaching consequences to the experimental design, replacing enthusiasm with disappointment. As illustrated with the discussion about the decisions over the experimental design (sample size, subject population, the selection of the treatment variable, its level and variation), generalizability and extrapolation can be jeopardized in the process due to the influence of political and financial values. However, this does not necessarily mean that objectivity would been compromised, because there are other possible aims that RFE's can serve. Rather it illustrates complex value trade-offs between epistemic and non-epistemic values that require attention and careful judgment of the researcher.

Moreover, studying the welfare impacts or distributional impacts, give researcher an opportunity to gather additional evidence, which presents an opportunity to increase the relevance and the quality of the study by adding value-laden theoretical concepts into the research agenda, as was done in BIE Finland. However, if these theoretical aspects are not dealt with appropriately, on integrated into the methodological core of the RFE, they can present epistemic risks too. Finally, the interpretation of the results, without additional evidence, can lead to speculative and subjective analysis of the experimental findings without deeper knowledge about the mechanisms of behaviour. This also suggests that the results of the RFE's to be useful in the first place, might require an additional evidence from subgroup analyses and welfare surveys that significantly lower the evidential standards of the EBP. However, if additional evidence on the secondary effects of the treatment are available, as is the case in BIE Finland, there is a risk of over- or underemphasizing the different pieces of evidence and therefore; the overall effects of the treatment.

As I have argued all these routes show how non-epistemic values can become constitutive values of the research, showing inadequacy of the value-free ideal to provide normative guidance in the evidence-based policymaking. I have argued that EBP should expand their views beyond narrow mechanical objectivity to cover more explicitly the constituents of the trained judgment both in the experimental design and analysis of the results. This requires an elaboration of the epistemic and non-epistemic values as well. The role of non-epistemic values in the evidence-based policy cannot be eliminated entirely. I have suggested that EBP could build on evolving literature and case studies that elaborate inductive risk and how values manifest when science and policy overlap.

References

Andreasen, R. & Doty, H. (2016): "Measuring Inequality: the role of values and inductive risk", in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Angrist, J. D. & Pischke, J. (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking Con out of Econometrics", *Journal of Economics Perspectives*, vol. 24, number 2, spring 2010, pp. 3-30.

Baldassarri, D. & Abscal, M. (2017): "Field Experiments Across the Social Sciences", *Annual Review of Sociology*, 2017, 43, pp. 41-73.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., Udry, C. (2015): "A multifaceted program causes lasting progress for the very poor: evidence from six countries" *Science 348*.

Banerjee, A., Chassang, S, & Snowberg, E. (2016): "Decision theoretic approaches to experiment design and external validity" *NBER Working Paper 22167*.

Banerjee, A. (2019): "Nobel Lecture: Abhijit Banerjee, Prize in Economic Sciences 2019", online source, available at https://www.youtube.com/watch?v=XvyMO7CmFlk (19.4.2020).

Banerjee, A., & Duflo, E. (2011): *Poor Economics: Barefoot Hedge-fund Managers, DIY Doctors and the Surprising Truth about Life on Less Than $1 a Day*, Penguin Books, London.

Banerjee, A., & Duflo, E. (2019): *Good Economics for Hard Times: Better Answers to Our Biggest Problems*, Allen Lane, Great Britain.

Barnow, B. and Greenberg, D. (2015): "Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence From Previous Social Experiments", *Evaluation Review*, 2015, Vol. 39(2), pp. 179-228.

Blalock, H. (1991): "Are There Really Any Constructive Alternatives to Causal Modeling?", *Sociological Methodology*, Vol. 21 (1991), pp. 325-335.

Bluhm, R. (2016): "Inductive risk and the Role of Values in Clinical Trials", in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Bossuroy, T. & Delavallade, C. (2016): "Experiments, policy, and theory in development economics: a response to Glenn Harrison's 'field experiments and methodological intolerance'", *Journal of Economic Methodology*, 23:2, pp. 147-156.

Boumans, M. (2016): "Methodological ignorance: A comment on field experiments and methodological intolerance", *Journal of Economic Methodology*, 23:2, pp. 139-146.

Brodeur, A., Cook, N., Heyes, A. (2018): "Methods Matter: P-Hacking and Causal Inference in Economics", *IZA Institute of Labor Economics Discussion Paper Series*, No. 11796, August 2018.

Business Standard (15.10.2019): online source, available at https://www.business-standard.com/article/current-affairs/in-backdrop-of-economics-nobel-announcement-letter-panning-rct-surfaces-119101500595_1.html#.Xf9oiuer-A4.twitter (19.4.2020).

Cairney, P. (2016): *The Politics of Evidence-Based Policy*, Palgrave Macmillan, London.

Campbell, D. (1969): "Reforms as Experiments", *American Psychologist* 24(4), pp. 409-429.

Card, D., DellaVigna, S., Malmendier, U. (2011): "The Role of Theory in Field Experiments", *The Journal of Economic Perspectives,* Vol. 25, No. 3, Summer 2011, pp. 39-62.

Cartwright, N. (2007): "Are RCTs the Gold Standard", *BioSocieties,* 2007 Vol. 2 (1), pp. 11-20.

The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel (2019): "Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019, Understanding development and poverty alleviations", The Royal Swedish Academy of Sciences, 2019.

Cook, T. (2014): "Generalizing Causal Knowledge in the Policy Sciences: External Validity as a Task of Both Multi-Attribute Representation and Multi-Attribute Extrapolation", *Journal of Policy Analysis and Management,* 33(2), pp. 527-536.

Currie A. (2015): "Philosophy of Science and the Curse of the Case Study", in: Daly C. (eds) *The Palgrave Handbook of Philosophical Methods,* Palgrave Macmillan, London.

Daston, L., and Galison, P. (1992): "The Image of Objectivity", *Representations,* No. 40, Special Issue: Seeing Science (Autumn, 1992), pp. 81-128.

Daston, L., and Galison, P. (2007): *Objectivity*, Zone Books, New York.

Deaton, A. (2010a): "Understanding the Mechanisms of Economic Development", *The Journal of Economic Perspectives,* Vol. 24, No. 3, Summer 2010, pp. 3-16.

Deaton, A. (2010b): "Instruments, Randomization, and Learning about Development", *Journal of Economic Literature,* Vol. 48, No. 2, June 2010, pp. 424-455.

Deaton, A. & Cartwright, N. (2017): "Understanding and misunderstanding randomized controlled trials", *Social Science & Medicine,* electronically available at *https://www.sciencedirect.com/science/article/pii/S0277953617307359?via%3Dihub*

De Wispelaere, J., Halmetoja, A., Pulkka, V. (2018): "The Rise (and Fall) of the Basic Income Experiment in Finland", *Focus,* vol. 19 3/2018, pp. 15-19.

Douglas, H. (2000): "Risk and Values in Science", *Philosophy of Science, Vol. 67, No. 4 (December 2000), pp. 559-579.*

Douglas, H. (2004): "The Irreducible Complexity of Objectivity", *Synthese* 138 (2004), pp. 453-473.

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

Douglas, H. (2016): "Foreword" in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Dolan, P., & Galizzi, M. (2014): "Getting policy-makers to listen to field experiments", *Oxford Review of Economic Policy,* Volume 30, Number 4, 2014, pp. 725-752.

Dreze, J. (2018): "Evidence, policy and politics: A commentary on Deaton and Cartwright", *Social Science & Medicine,* 210 (2018), pp. 45-47.

Duflo, E., Dupas, P., Kremer, M. (2006): "Education and HIV/AIDS prevention: Evidence from a randomized evaluation in Western Kenya", *World Bank Policy Research Working Paper* 4024, October 2006.

Duflo, E., & Kremer, M. (2005): "Use of Randomization in the Evaluation of Development Effectiveness", in Pitman, George K., Feinstein, O., and Ingram, G. (eds.) *Evaluating Development Effectiveness: World Bank Series on Evaluation and Development Volume 7*, 2005, Transaction Publishers, New Brunswick, New Jersey.

Duflo, E. (2017):, "The Economist as Plumber", *American Economic Review: Papers & Proceedings 2017*, 107(5), pp. 1-26.

Duflo, E. (2019): "Nobel Lecture: Esther Duflo, Prize in Economic Sciences 2019", online source, available at https://www.youtube.com/watch?v=KFRnY-5K5OU (19.4.2020).

Elliott, K. & Richards, T. (2016): "Introduction" in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Faverau, J. (2016): "On the analogy between field experiments in economics and clinical trials in medicine", *Journal of Economic Methodology*, 23:2, pp. 203-222.

Faverau, J. & Nagatsu M. (2020a): "Two Strands of Field Experiments in Economics: A Historical-Methodological Analysis", *Philosophy of the Social Sciences*, 50(1), pp. 45-77.

Faverau, J. & Nagatsu, M. (2020b): "Holding back from theory: limits and methodological alternatives of randomized field experiments in development economics", *Journal of Economic Methodology*, January 2020, DOI: 10.1080/1350178X.2020.1717585.

Ferber, R., & Hirsch, W. (1978) "Social Experimentation and Economic Policy: A Survey", *Journal of Economic Literature*, Vol. 16, No. 4, December 1978, pp. 1379-1414.

Foreign Policy (22.10.2019): available at https://foreignpolicy.com/2019/10/22/economics-development-rcts-esther-duflo-abhijit-banerjee-michael-kremer-nobel/ (19.4.2020).

Galison, P. (2015) "The Journalist, the Scientist, and Objectivity", in Flavia Padovani, Alan Richardson, Jonathan Y. Tsou (eds.), *Objectivity in Science: New Perspectives from Science and Technology Studies*, Boston Studies in the Philosophy and History of Science 310, Springer International Publishing Switzerland.

Gerber, A. & Green, D. (2002): "Reclaiming the Experimental Tradition in Political Science" in *Political Science: State of the Discipline* edited by Katznelson, I. & Milner, H., New York, W.W. Norton, pp. 805-832.

Greenberg, D, Barnow, B. (2014), "Flaws in Evaluations of Social Programs: Illustrations From Randomized Controlled Trials", *Evaluation Review*, 2014, Vol. 38(5), pp. 359-387.

Greenberg, D. & Robins (1986): "The Changing Role of Social Experiments in Policy Analysis", *Journal of Policy Analysis and Management*, Winter 1986, 5, 2, p. 340.

Greenberg, D.; Shroder, M.; Onstott, M. (1999): "The Social Experiment Market", *Journal of Economic Perspectives*, Vol. 13, Number 3, Summer 1999, pp.157-172.

Greenberg, D. & Shroder, (2004): *The Digest of Social Experiments*. The Urban Institute Press. Washington.

Grüne-Yanoff, Till (2012): “Old wine in new casks: libertarian paternalism still violates liberal principles”, *Social Choice and Welfare* 38, pp. 635-645.

Guala, F. (2005): *The Methodology of Experimental Economics*, Cambridge University Press, New York.

The Guardian (14.10.2019): available at https://www.theguardian.com/world/2019/oct/14/economics-nobel-prize-abhijit-banerjee-esther-duflo-michael-kremer (19.4.2020).

Haavelmo, T. (1944): “The probability approach in econometrics”, Supplement to *Econometrica*, 12, pp. 1-115.

Hacking, I. (2015): ”Let’s Not Talk About Objectivity”, in Flavia Padovani, Alan Richardson, Jonathan Y. Tsou (eds.), *Objectivity in Science: New Perspectives from Science and Technology Studies*, Boston Studies in the Philosophy and History of Science 310, Springer International Publishing Switzerland.

Hallituksen julkaisusarja 10/2015: ”Pääministeri Juha Sipilän hallituksen strateginen ohjelma 29.5.2015”, https://valtioneuvosto.fi/documents/10184/1427398/Ratkaisujen+Suomi_FI_YHDISTETTY_netti.pdf (28.1.2019).

Harrison, G., & List, J. (2004): “Field Experiments”, *Journal of Economic Literature*, Vo. 42, No. 4, December 2004, pp. 1009-1055.

Harrison, G. (2013): “Field experiments and methodological intolerance”, *Journal of Economic Methodology*, 20:2, pp. 103-117.

Harrison, G. (2014): “Cautionary notes on the use of field experiments to address policy issues”, *Oxford Review of Economic Policy*, Volume 30, Number 4, 2014, pp. 753-763.

Hausman, J. & Wise, D. (1985): “Introduction to ‘Social Experimentation’” in Jerry A. Hausman and David A. Wise (eds.) *Social Experimentation*, University of Chicago Press.

Heckman, J. (2019): “Randomization and Social Policy Evaluation Revisited”, *HCEO Working Paper Series*, Working Paper 2019-073, 12/2019.

Hempel, C. (1965): “Science and Human Values.” In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, by Hempel, C., pp. 81– 96. New York: Free Press.

Hiilamo, H. (8.2.2019): online source, available at https://www.helsinki.fi/en/news/nordic-welfare-news/heikki-hiilamo-disappointing-results-from-the-finnish-basic-income-experiment

Hill, A., (1965): “The environment and disease: Association or causation?” *Journal of the Royal Society of Medicine*, 58 (5), pp. 295–300.

Hämäläinen, K.; Kanninen, O.; Simanainen, M.; Verho, J. (2019): ”Perustulokokeilun ensimmäinen vuosi”, *VATT muistiot 56*, Helsinki.

Jacobin (12.1.2019): available at https://www.jacobinmag.com/2019/12/basic-income-finland-experiment-kela (19.4.2020).

James, W. (1896): “The Will to Believe.” *The New World* 5:327– 47.

J-PAL (2020): online source, available at https://www.povertyactionlab.org/evaluations (19.4.2020).

Kangas, O., Simanainen, M., Honkanen, P. (2017): "Basic Income Experiment in the Finnish Context", *Intereconomics,* vol. 52, number 2, pp. 87-91.

Kangas, O. (2016): online source, available at https://tutkimusblogi.kela.fi/arkisto/3316 (19.4.2020).

Kangas, O., Jauhiainen, S., Simanainen, M., Ylikännö M. (2019): "The Basic Income Experiment 2017-2018 in Finland. Preliminary results", *Reports and Memorandums of the Ministry of Social Affairs and Health,* 2019:9.

Keane, M. (2010a): "Structural vs. atheoretic approaches to econometrics", *Journal of Econometrics,* 156 (2010), pp. 3-20.

Keane, M. (2010b): "A Structural Perspective on the Experimentalist School", *The Journal of Economic Perspectives,* Vol. 24, No. 2, Spring 2010, pp. 47-58.

Kela (The Social Insurance Institution) (2016): "From idea to experiment: Report on universal basic income experiment in Finland", *Kela working papers 106*, Helsinki.

Khosrowi, D. (2018): "Trade-offs between epistemic and moral values in evidence-based policy", *Economics and Philosophy*, Cambridge University Press, electronically available at https://www.cambridge.org/core/journals/economics-and-philosophy/article/tradeoffs-between-epistemic-and-moral-values-in-evidencebased-policy/DC1EC271584CD73D9728BDEF684C1977

Khosrowi, D. & Reiss, J., (2019): "Evidence-Based Policy: The Tension Between the Epistemic and the Normative", *Critical Review,* 31:2, pp. 179-197.

Kuhn, T., 1962 [1970], *The Structure of Scientific Revolutions*, Second edition, Chicago: University of Chicago Press.

Koskinen, I., (2018): "Defending a Risk Account of Scientific Objectivity", *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy053

La Caze, A. & Colyvan, M. (2017): "A Challenge for Evidence-Based Policy", *Axiomathes,* 2017, 27, pp. 1-13.

Leamer, E. (1983) "Let's Take the Con Out of Econometrics", *American Economic Review,* 73(1): 31–43.

Leamer, E. (2010): "Tantalus on the Road to Asymptopia", *The Journal of Economic Perspectives,* Vol. 24, No. 2, Spring 2010, pp. 31-46.

Levitt, S. & List, J. (2009): "Field Experiments in Economics: The Past, The Present, and The Future", *NBER Working Paper No. 14356,* September 2008.

List, J. (2011): "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off", *Journal of Economic Perspectives,* Volume 25, Number 3, Summer, 2011, pp. 3-16.

Longino, H., (1990): *Science as social knowledge: values and objectivity in scientific inquiry*, Princeton University Press, Princeton, New Jersey.

Ludwig, J., Kling, J., Mullainathan, S. (2011): ”Mechanism Experiments and Policy Evaluations”, *NBER working paper series, Working Paper 17062*.

Morton, R. & Williams, K. (2010): *Experimental Political Science and the Study of Causality: From Nature to the Lab,* Cambridge University Press, New York.

Nagel, E. (1961): *The Structure of Science, Problems in the Logic of Scientific Explanation,* Routledge, London.

The New York Times (20.7.2017): available at https://www.nytimes.com/2017/07/20/opinion/finland-universal-basic-income.html.

Niiniluoto, I. (1993): “The aim and structure of applied research”, *Erkenntnis* 38: pp. 1-21.

OECD, Development Assistance Committee (2018): “OECD DAC Evaluation Criteria: Summary of consultation responses 2018”, available at https://ieg.worldbankgroup.org/sites/default/files/Data/DAC-Criteria/ConsultationReport_EvaluationCriteria.pdf.

OECD, Development Assistance Committee (2019): online source, available at https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm (19.4.2020).

Orr, L. (1999): *Social experiments: Evaluating public programs with experimental methods,* SAGE Publications, California.

Plott, C. (1986): “Laboratory Experiments in Economics: The implications of posted-price institutions, *Science*, 232:732-38.

Plott, C. (1997): “Laboratory Experimental Testbeds: Application to the PCS Auction”, *Journal of Economics & Management Strategy,* Volume 6, Number 3, Fall 1997, pp. 605-638.

Plutynsky, A. (2016): “Safe or Sorry? Cancer Screening and Inductive Risk”, in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Rawlings, L. (2005): “Operational Reflections on Evaluating Development Programs”, in Pitman, G., Feinstein, O., Ingram, G. (eds.) *Evaluating Development Effectiveness: World Bank Series on Evaluation and Development Volume 7*, 2005, Transaction Publishers, New Brunswick, New Jersey.

Politiikasta (3.3.2018): available at https://politiikasta.fi/perustulokokeilu-ei-johda-perustulon-toteuttamiseen/ (19.4.2020).

Reiss, J. & Sprenger, J. (2017): "Scientific Objectivity", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/> (19.4.2020).

Reiss, J. (2013): *Philosophy of Economics: A Contemporary Introduction,* Routledge, New York.

Reiss, J. (2014): “Struggling over the Soul of Economics: Objectivity versus Expertise” in *Experts and Consensus in Social Sciences,* by Martini C. & Boumans, M. (eds.), Springer.

Reiss, J., (2018): “Against external validity”, *Synthese* 196, pp. 3103-3121.

Risjord, M. (2014): *Philosophy of Social Science: A Contemporary Introduction*, Routledge, New York.

Ross, D. (2013): "Introduction to discussion forum on Glenn W. Harrison's 'field experiments and methodological intolerance'", *Journal of Economic Methodology*, 23:2, pp. 127-129.

Roth, A. (1986): "Laboratory Experimentation in Economics", *Economics and Philosophy*, 2, 1986, pp. 245-273.

Roth, A. (1988): "Laboratory Experimentation in economics: A methodological overview", *Economic Journal*, 98:974-1031.

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

Ruphy, S., (2006): "'Empirism all the way down': a defense of the value-neutrality of science in response to Helen Longino's contextual empirism", *Perspectives on Science*, 2006, vol. 14, no. 2.

Ruzzene, A. (2015): "Policy-making in developing countries: from prediction to planning", *Journal of Economic Methodology*, 22:3, pp. 264-279.

Santos, A. (2012): "The facts and values of experimental economics", in *Facts, Values, and Objectivity in Economics* by Caldos, J., Neves, V. (eds.)., Routledge, New York.

Smith, K., & Joyce, K. (2012): "Capturing complex realities: Understanding efforts to achieve evidence-based policy and practice in public health", *Evidence & Policy*, 8 (1), 57–78.

de Souza Leao, L. & Eyal, G. (2019): "The rise of randomized controlled trials (RCTs) in international development in historical perspective", *Theory and Society*, 2019, 48, pp. 383-418.

Stanev, R. (2016): "Inductive Risk and Values in Composite Outcome Measure", in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, K. & Richards T. (eds.), Oxford University Press, 2017.

Stegenga, J. (2016): "Drug Regulation and the Inductive Risk Calculus", in *Exploring Inductive Risk: Case Studies of Values in Science* by Elliott, Kevin C. and Richards Ted (eds.), Oxford University Press, 2017.

Svorencik, A. (2015): "The Experimental Turn in Economics: A History of Experimental Economics", University of Utrecht: Utrecht School of Economics Dissertation Series #29, electronically available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2560026

Tsou, J., Richardson, A., Padovani, F. (2015): "Introduction: Objectivity in Science" in Flavia Padovani, Alan Richardson, Jonathan Y. Tsou (eds.), *Objectivity in Science: New Perspectives from Science and Technology Studies*, Boston Studies in the Philosophy and History of Science 310, Springer International Publishing Switzerland.

Valtioneuvoston kanslia (2016): "Ideasta kokeiluun: esiselvitys perustulokokeilun vaihotehdoista", *Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 13/2016*, edited by Kangas, O. & Pulkka, V., available at https://tietokayttoon.fi/documents/10616/2009122/13-2016_Ideasta+kokeiluun.pdf/c758c343-2687-4dea-869e-5dbdb14e888f/13-2016_Ideasta+kokeiluun.pdf?version=1.0 (19.4.2020).

Valtioneuvoston päätös VNK/2019/111: ”Valtioneuvoston päätös strategisen tutkimuksen teema-alueista ja painopisteistä vuodelle 2020, Valtioneuvoston yleisistunto 10.10.2019”, available at https://www.aka.fi/globalassets/33stn/teemat/20191010_vn_teemapaatos.pdf (22.3.2020).

Weber, M. (1917/1988): “Der Sinn der ‘Wertfreiheit’ der soziologischen und ökonomischen Wissenschaften”, reprinted in Gesammelte Aufsätze zur Wissenschaftslehre, Tübingen: UTB, 451–502.

White, H. (2005): “Challenges in Evaluating Development Effectiveness”, in Pitman, G., Feinstein, O., Ingram, G. (eds.) *Evaluating Development Effectiveness: World Bank Series on Evaluation and Development Volume 7*, 2005, Transaction Publishers, New Brunswick, New Jersey.

Wilcox, N. (2016): “Robert A. Millikan meets the credibility revolution: comment on Harrison (2013), ‘field experiments and methodological intolerance’”, *Journal of Economic Methodology*, 23:2, pp. 130-138.

Wilholt, T. (2009): “Bias and Values in Scientific Research.” *Studies in History and Philosophy of Science*, Part A 40(1): 92– 101.

The World Bank (2005): *The logframe handbook: a logical framework approach to project cycle management, Working paper 31240,* electronically available at http://documents.worldbank.org/curated/en/783001468134383368/The-logframe-handbook-a-logical-framework-approach-to-project-cycle-management.

Yle (12.5.2017): available at https://yle.fi/uutiset/3-9607245 (19.4.2020).

Yle News (2.4.2019): available at https://yle.fi/uutiset/osasto/news/finlands_basic_income_trial_did_not_make_recipients_passive_govt_research_finds/10718492 (19.4.2020).