Erasmus School of philosophy[1]

Bachelor Thesis philosophy of a specific discipline

Transcending experimental boundaries

A philosophical analysis of the necessity of Transcendental Realism for achieving external validity in the Credibility Revolution

July 17, 2023

Author:
Jelmer de Jong
Student number:
484379
Main study:
Econometrics and Operations Research

Supervisor:
Prof.dr. JJ Vromen
Advisor:
Prof.dr. FA Muller
Number of Words:
8642

Abstract

Identifying causal effects in economics is a challenging endeavour. Therefore, economists initiated a credibility revolution which attempts to create methods that truly identify the causal effect via the use of quasi-experimental methods. However, their methods receive criticism for lacking external validity. This paper considers the philosophical framework necessary for obtaining external validity. While positivism shares similarities with the approach in the credibility revolution, it is discarded due to its superficiality. Consequently, I argue for the necessity of accepting transcendental realism to obtain external validity in the credibility revolution. I show that adopting this framework facilitates the logical possibility of external validity as well as its methodological solutions.

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of philosophy or Erasmus University Rotterdam.

Table of contents

1 List of acronyms

CR Credibility Revolution 3

DiD Difference in Difference 8

IV Instrumental Variables 8

RCT Randomised Controlled Trials 6

RDD Regression Discontinuity Design 8

RBC Rubin Causal Model 6

TR Transcendental Realism 4

2 Introduction

Econometricians encounter a particular challenge due to the inherent nature of the subject matter: the economy, which essentially evolves around human beings guided by their own volition. This challenge contrasts the natural sciences, where research materials conform to regular and predictable laws. One domain where this contrast becomes apparent is the domain of causal inference, where the goal is to uncover causal relations between phenomena. While the natural sciences can identify true causal effects, econometricians struggle to distinguish between causation and correlation. Thus, it is no surprise that since the beginning of the discipline, econometrics has met with much criticism. John Maynard Keynes dubbed it as statistical alchemy, and Edward E. Leamer insisted on taking the "con" out of econometrics. To counter this critique, economists initiated developments that led to new methods in econometric methodology, which have resulted in the so-called Credibility Revolution (CR). The aim of the CR is to attempt to increase the credibility of empirical results, i.e. results obtained from empirical research (Angrist and Pischke, 2010). The new methods use a (quasi)-experimental approach to identify a causal effect, thereby successfully establishing the internal validity of their results. One can interpret internal validity as whether the causal effect identified through empirical investigation is, in fact, the true causal effect (Campbell, 1957). However, critics argue that these quasi-experimental methods lack external validity, which one interprets as the validity of generalising the results to contexts other than the research context (Deaton, 2009; Urzua et al., 2009; Rosenzweig and Wolpin, 2000; Heckman, 1997). Since researchers use econometrics to estimate the effect of policies or predict what effect a policy will have, it is impossible to accurately predict a policy's effect without external validity, as it is implausible that the research context is the same as the context of the policy. Therefore, researchers should obtain internal and external validity of their results to ensure well-constructed policies. In light of econometrics' longstanding criticisms and challenges, this paper aims to explore the philosophical foundations of external validity within the context of the CR. Through this exploration, this paper aims to contribute to the ongoing discourse on causal inference and its implications for decision-making processes by critically examining the complexities inherent in econometric modelling.

Current research regarding external validity focuses mainly on methodological solutions where the line between internal and external validity is blurred (Bertanha and Imbens, 2020; Wing and Bello-Gomez, 2018; De la Cuesta et al., 2022). However, they need to consider all possible influences, hence these studies do not do justice to the inherent complexities of the economy. Furthermore, external validity amounts to the question of the validity of scientific statements. A philosophical analysis must determine on what grounds a scientific statement is

considered meaningful or truthful to establish this validity. Similar to not taking into account all possible influences, recent philosophical research also neglects external validity through the specific setup of their experimental methods (Reiss, 2013; Guala, 1999, 2003; Guala and Mittone, 2005; Bardsley, 2010).

This paper aims to contribute to the existing literature by providing broader philosophical foundations for external validity concerning the quasi-experimental methods in the CR. To address the problem of external validity within the CR, it is essential to consider philosophical frameworks that offer insights into the essence of scientific statements and the challenges faced by econometricians. This framework should facilitate the possible methodological solutions for external validity.

In this paper, I argue that embracing a Transcendental Realism (TR) framework is necessary for achieving external validity within the CR. Positivism has been the predominant methodological framework in economics since Milton Friedman (Friedman, 1953). At first glance, the CR seems to align well with positivism. However, achieving external validity proves to be challenging as positivism only focuses on explaining empirical economic occurrences and disregards the unobservable underlying phenomena. The numerous approaches to acquiring external validity depend on looking further than what is observable, thereby recognising the importance of contextual circumstances for obtaining external validity. Therefore, a TR framework should be accepted. TR, argued for by Lawson in his book *Economics & Reality* (Lawson, 1997, Henceforth as TL), recognises the importance of unobservable phenomena, the social and contextual circumstances. Moreover, Lawson argues the world consists of observations of actual events and unobservable phenomena such as structures, mechanisms and tendencies. These concepts can be interpreted respectively as the enduring patterns that shape social phenomena, the link between these phenomena and the inclination of these structures and mechanisms. Economic examples of unobservable phenomena are cultural values that shape economic behaviour or beliefs and attitudes that influence investment behaviour. While the effects of these variables might show themselves to us, they remain unobservable. Accepting TR opens the possibility of attaining empirical regularities akin to the natural sciences and facilitates viable methodological solutions to address the challenge of external validity. Lawson defines empirical regularities as a ubiquitous phenomenon of the type, whenever event x then event y (TL, p.19).

The paper and its corresponding arguments are structured as follows: Section 3 introduces the CR. This section consists of three subsections. The first subsection, 3.1, comprises the appropriate methods and strategies for identifying causal effects. Subsequently, the concept of external validity, in light of the CR, is discussed in section 4. In the final subsection, 4.1,

possible solutions for the problem of external validity are considered. Following this, section 5 explains the framework of logical positivism and its verification principle. It concludes with a deliberation on whether it might be considered for the problem of external validity. In the penultimate section 6, the framework of TR is explained extensively. This is followed by the main arguments, arguing for the necessity of accepting TR to achieve external validity. Finally, in section 7, the main concepts and arguments are summarised and concluded.

3 Credibility Revolution

Each year the economics variant of the Nobel prize, The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, is awarded for significant contributions to the discipline of economics. Both the Nobel Prizes in 2019 and 2021 were awarded to economists contributing to a revolution in economics. Economists Guido Imbens, David Card and Joshua Angrist received the prize in 2021 for their contributions to causal inference analysis in economics. Their contributions form the basis for the CR in economics. Esther Duflo, Michael Kremer and Abhijit Banerjee received the Nobel Prize in 2019 for their work applying experimental methods to alleviate global poverty. Their work also significantly influenced the CR. The award of two prizes already signals the influence of the CR on economics. To explain and define these methods, I first briefly overview causal inference and its problems in economics.

3.1 Identifying causal effects

The goal of causal inference in economics is to study causal relations between economic variables. Philosophers have long considered causality, of which famous examples are Kant, Hume and Aristotle. This paper defines causation as an unidirectional influence of a variable, the cause, on another variable; the change in the state of the affected variable is called the effect. Examples from economics are the effect of a price increase of chocolate on the demand for chocolate, the effect of education on wages or the effect of an increase in the interest rate on consumption. Uncovering and quantifying these effects is crucial to inform and design economic policy or to decide on business strategies. However, researchers struggle to identify a causal effect accurately. Incorrect identification of the causal effect will, in turn, lead to misinformation of policy questions. The difficulty in finding these causal relations is to distinguish between correlation and causation. When two variables are correlated, there exists a statistical relation between them, i.e. they are linearly related in the sense that when one variable increases, the other increases as well. On the other hand, this does not have to mean that the other variable causes an increase

in the other variable. Hence the notorious maxim: correlation does not imply causation.

To give an example of the confusion between causation and correlation, consider the effect education has on wages. Suppose there is a dataset showing that people with high wages typically receive more years of education. An inexperienced researcher might infer that receiving more education is the cause of earning higher wages later on. On the other hand, high intelligence is also more prominent among well-educated high-earners. Therefore, it is difficult to distinguish in which direction the causality runs. Does education cause high wages, or does the workers' intelligence cause high wages? The difficulty arises from the fact that the economy is a hugely complex system where countless variables interact. Thus, when the goal is to identify a causal relation between two variables, one needs to control for the other interacting variables. The natural sciences utilise experiment control to account for interacting factors. The CR copies this idea and encourages researchers to adopt quasi-experimental methods. I explain these methods using the Rubin Causal Model (RBC) (Rubin, 1974).

3.2 Rubin Causal Model

The RBC is a model to analyse the theory of cause and effect statistically. Rubin takes the experimental methods of Randomised Controlled Trials (RCT) as a foundation. These methods are experimental methods where subjects are divided into two groups, the first group, which we call T, will receive the treatment, and the second group, C, will not receive the treatment and is named the control group. The basic idea of an RCT is to control the variables which are not under investigation and take advantage of the random treatment allocation. The origins of RCT are in medicine, where they are applied to discover the causes of unrevealed and unpredictable effects of medical treatments (Meldrum, 2000).

Let us look at the RBC with the help of an example from economics. Suppose the government implements an extensive job training programme for a specific population, such as unemployed immigrants. Half of this specific population is assigned to T and will receive more specialised training on finding a job; the other half is assigned to the control group and receives regular job training. Let variable Y denote whether or not the subject finds a job after one year. In this case, the outcome is dichotomous; however, it could also be continuous or multinomial. To measure the causal effect of the treatment, you would ideally want to measure the difference between the outcome effects of someone receiving the treatment and someone not. If we let $y_i(T)$ and $y_i(C)$ be the potential outcome for subject i in the treatment group and the control group, respectively, the causal effect of the treatment is identified as $y_i(T) - y_i(C)$. However, it is impossible for a subject to both receive and not receive the treatment simultaneously. Hence

the causal effect cannot be identified. This identification problem is called the fundamental problem of causal inference (Holland, 1986).

RCTs provide the following solution. Suppose there are two subjects; one is in the treatment group and the other in the control group. The causal effect of the treatment is then defined as $\frac{1}{2}\big[y_1(T) - y_1(C) + y_2(T) - y_2(C)\big]$ (Rubin, 1974). However, we do not observe both individual outcomes and can only provide an estimate using $y_1(T) - y_2(C)$ or $y_2(T) - y_1(C)$, depending on which subject is assigned to which group. At this point, the randomised element of RCT becomes important. Rubin (1974) proposes randomly assigning subjects in the target population to either the treatment or control groups. Consequently, there is an equal chance for either subject to be assigned to the treatment or the control group. The average causal effect of the observed outcomes then is $\frac{1}{2}\big[y_1(T) - y_2(C)\big] + \frac{1}{2}\big[y_2(T) - y_1(C)\big]$, which is equal to the initial specified causal effect of the treatment. Thus, randomisation provides the possibility to obtain an unbiased estimate of the causal effect of the treatment.[2].

These methods provide great possibilities for identifying and estimating the causal effect of a treatment in an RCT. The application of these methods has been widespread in economics but is prevalent in the field of development economics (Banerjee et al., 2016), which led to the award of the Nobel Prize in economics, as discussed earlier. However, there is a limit to what RCTs can achieve. For example, suppose a researcher designs an RCT to discover the effect of a basic income for two years. In that case, this will cost enormous amounts of money, and questions arise whether such an experiment is ethically justifiable as the researcher allocates the basic income to only a set of individuals. Thus, there are significant ethical and practical problems with RCTs in economics. To combat these problems, researchers resort to studying experimental situations where natural forces or human organisations might be of help. These situations provide possibilities for quasi or natural experiments, which are the main approaches in the CR.

The methods in the CR are not merely a reaction to the rise of the sometimes impractical RCTs; they result from critical reflections on econometrics. The culmination of this critique comes from a paper which proposes to take the con out of econometrics by Leamer (1983). In his influential paper, Leamer argues that economics is too dependent on econometric methods, which lack robustness, are too sensitive to model specification, and cannot distinguish correlation and causation. He argues for adopting a more holistic approach by integrating multiple methods and extensive sensitivity analysis. Angrist and Pischke (2010) argue that improving empirical research design can take the con out of econometrics. They also highlight the importance of the

[2] An unbiased estimate means that the expected value of an estimator is equal to the true value of the estimated value, i.e. $E(Y) = Y$

vast increase in available data, the fact that econometric methods improved significantly and the transparency surrounding these new research designs.

3.3 Quasi-experimental methods

The significant developments in research design are the leading actor in the CR. These new designs avoid the fallacies made when designing RCTs. The most commonly used quasi-experimental methods are Instrumental Variables (IV), Regression Discontinuity Design (RDD), and Difference in Difference (DiD). I explain these methods shortly.

IV is a quasi-experimental method that, as its main attribute, tries to solve the endogeneity problem in causal inference. Endogeneity occurs as a result of omitted variable bias, simultaneity or measurement error. If present, endogeneity induces a correlation between an explanatory variable and the error term in a regression, i.e., a non-observed variable or confounding variable influences both the explanatory variables and the outcome. Endogeneity produces a bias in the estimated treatment variable. An example of endogeneity is already briefly mentioned when discussing the effect of education on wages. For example, intelligence might be a confounding variable as it influences both the education level and plausibly influences the wage. Thus, the causal effect is misidentified when this variable is not present in the regression. An IV approach solves the problem of endogeneity as follows. First, the researcher finds a variable that is not correlated with the error term or any other of the explanatory variables but is correlated with the explanatory variable responsible for the endogeneity and the treatment outcome. However, note that the correlation between the IV and the treatment outcome happens through the explanatory variable. In other words, an increase in the IV increases the explanatory variable, subsequently increasing the treatment outcome. Next, the researcher performs two regressions: the endogenous explanatory variables are regressed on the IV, and the treatment outcome is regressed on the fitted values of the first regression. This approach is called two-stage least squares.[3] Card (1993) proposes to use college proximity as an IV for estimating and identifying the causal effect of education on wage. The rationale is that college proximity increases the education rate but does not influence wages directly.

RDD is a quasi-experimental method that uses a specific assignment mechanism to identify the causal effect. A common problem in non-random experimental designs is selection bias which arises when distorted selection rules are used as an assignment mechanism (Heckman, 1990). An example is the study of the effect of subsidies on company performance. When the better-known companies or companies which excel in lobbying receive this subsidy, the effect of

[3]For the technical details regarding this method, I refer the reader to chapter 5.7 of the book by Heij et al. (2004).

subsidies might be overrated as the well-performing companies have more chance of receiving the subsidy; this is selection bias. RDD studies use a running variable that serves as an indicator variable that determines whether an individual is assigned to the treatment group or the control group. The running variable, commonly, is continuous and is used as a threshold to allocate subjects into either the treatment or control group, depending on which side of the threshold they fall. Regarding the study into the effects of subsidies on company performance, one might use firm size as a running variable. The comparison of subjects around the threshold ensures the similarities between these subjects. Researchers assume that the subjects could have equally as well fallen on the other side of the threshold; hence randomisation is achieved. Consequently, with the help of the RBC, one can identify the causal effect of the treatment, in this case, the subsidy, on company performance.

DiD is a quasi-experimental method which requires some sort of natural experiment. The idea is that there is an exogenous source of variation which determines whether specific individuals receive the treatment. The researcher analyses the control and treatment groups at two points in time. Note that the control group must not experience any changes. However, the treatment group should receive treatment at one point in time. An example of a treatment in a natural experiment is the first DiD application to find the actual cause of cholera epidemics in London (Snow, 1856). Researchers first identified the contamination via air as the culprit. However, Snow argued that contaminated water was the cause. He compared two neighbourhoods over time, of which one received uncontaminated water in the second period. For the next step, the researcher calculated the corresponding estimator by subtracting the average gain over time in the treatment group from the average gain over time in the control group. Regarding the cholera experiment, Snow found that the number of infections decreased in the neighbourhood that received uncontaminated water in the second period. DiD is now one of the most used and best functional research designs in experimental economics (Angrist and Pischke, 2010). Moreover, it can be applied fruitfully to measure the effects of policy changes, provided that it concerns an exogenous policy change.

4 External validity

The common factor in the research designs of the CR is that they can identify causal relations without relying on impractical and expensive experiments. Consequently, these methods are able to attain internal validity, i.e. the true causal effect is identified. Obtaining internal validity is the goal of quasi-experimental methods in the CR. Although internal validity depends on

specific assumptions for each method, the CRs methods significantly improve economic practice. However, the CR receives two significant critiques. The first one is that the methods applied predominantly address questions in microeconomics, subsequently leaving the big economic questions unanswered. The second one is that the focus on internal validity has resulted in a lack of external validity. Angrist and Pischke (2010) summarise the critique, saying that their research designs have become too small and idiosyncratic. Among the critics are Deaton (2009); Urzua et al. (2009); Rosenzweig and Wolpin (2000); Heckman (1997), of which Deaton is the most prolific. He argues that results generated by quasi-experimental methods, albeit correctly identified, only help to inform policy if the researcher answers the question of why the effect happened. Moreover, he notes that policy changes do not reflect the experimental framework, i.e. researchers commonly generalise results under the notion that the experiment is like the policy. According to Deaton, the problem with this is that unobservable factors are not considered as the "like" is merely based on observable similarities. Finally, he argues for quasi-experimental methods to become more theory-driven.

Notions of external and internal validity originated more than 50 years ago (Campbell, 1957). External validity concerns generalisation; specifically, it concerns whether a causal statement still holds over variations of subjects, populations, time, different settings or treatments. Previously, external validity also contained the concept of construct validity, which concerns to what extent the particular treatments, subjects, and observations are generalisable to the theoretical concepts they are supposed to measure. However, these two types of validity are now demarcated mainly on pragmatic reasons as both the concepts receive such a critique that a division was necessary (Cook and Campbell, 1979). As critics criticise the CR on external validity, this paper does not go further into construct validity. The philosophical literature concerning external validity has not received much attention, and if it did, it primarily focuses on experimental economics performed in laboratory conditions (Reiss, 2013; Guala, 1999, 2003; Guala and Mittone, 2005; Bardsley, 2010). However, this problem is more comprehensive, as before one thinks about generalising causal evidence to, for example, other countries, one first has to overcome the extrapolation problem from laboratory to the real economy. Therefore, this extensive extrapolation will not be discussed further.

A lack of external validity poses problems for the practical applicability of the quasi-experimental methods. These methods are often applied to estimate the effect of a policy. For example, the natural experimental study examining the effect of a rise in the minimum wage on employment in two American states received much attention from media and politics (Card and Krueger, 1993). They find that an increase in the minimum wage does not lead to a decrease in the

employment rate. The paper received this attention mainly because the study obtained a result opposite conventional studies and theories. After reading about the study, a naive non-American politician might immediately advocate the rise of minimum wages. Notwithstanding the internal validity of this study, it is not plausible to generalise the result to, say, for example, Bulgaria, as the contextual differences will affect the causal effect as well. However, if some degree of external validity holds, the politician, depending on his goals, is vindicated in his advocacy for an increase in the minimum wage. Therefore, without some degree of external validity, the methods from the CR merely provide contingent causal relations.

4.1 Obtaining external validity

In this section, I discuss possible solutions for the problem of external validity. These solutions are found in contemporary literature and mainly concern qualitative or structural methodological solutions. While there are quantitative solutions, these solutions focus on specific research designs, whereas in this study, I consider general solutions for external validity applicable to all quasi-experimental research designs.[4] I opt to prioritise the general solutions as specific solutions contribute only to solving the problem of external validity for specific methods and do not counter the general critique received by the CR. Moreover, the technical solutions do not consider the complexity of the economy, which requires a conceptual investigation into the philosophical underpinnings.

Angrist and Pischke (2010) propose to combat external validity by doing more research on the same causal effect. Finding the same causal effect across multiple studies will increase confidence in the generalisability of the result. However, they concede that experimental results are always local and generalisation to external contexts will always be some sort of speculation. I agree with them and emphasise that the proposed solutions do not attempt or claim to solve the problem of external validity entirely. Experimentally established causal relations in economics will always depend on contextual factors. Nonetheless, by investigating these contextual factors, we can strive to lay the foundations for experimental research to achieve as much external validity as possible.

Banerjee et al. (2017) propose that experimental researchers should speculate in a structural manner as to whether their results have any external validity. This speculation should be precise and falsifiable such that researchers are encouraged to test these specific statements and conditions other than the original experiment. The experimenters should include in their

[4]For specific technical solutions see: Bertanha and Imbens (2020); De la Cuesta et al. (2022); Wing and Bello-Gomez (2018)

speculation: intuitive explanations of the present mechanisms, how inherent heterogeneity affects the results and how unobservable intuitions correlate with observable statistics.

On the other hand, Laurent (2000) formulates the problem of external validity as providing a guarantee that the structure of a causal model is similar to the structure of the whole system. For this approach, he builds on the work of Forrester (1997), who suggests creating a verbal model before building a mathematical one. This verbal model includes unobservable, not yet observed variables and important factors. Finally, Laurent (2000) argues a 3-step approach to modelling, 'proto modelling': in which qualitative research is performed to obtain the important variables and relationships, 'peri modelling', where the researcher ensures a correct functional form of the model and 'after modelling' to verify whether the structure and behaviour of the model are reliable.

Another potential solution is the method of mechanism mapping as proposed by (Williams, 2020). Problems with external validity arise due to an engagement of the contextual circumstances and how the policy changes these circumstances. Williams stresses the multidimensional character of the contexts, which requires a framework capable of handling this data. This framework should be capable of handling hard empirical evidence as well as soft unobservable evidence. Mechanism mapping is an approach consisting of three steps. In the first step, the theory of change is investigated, i.e. through what mechanism the policy had its effect in the original context. The second step involves uncovering all the contextual assumptions that lay at the foundation of the causal chain identified in the first step. The policy change will not have the intended effect if these assumptions do not hold in a specific context. Finally, in the third step, the context is characterised for each step in the causal mechanism such that discrepancies between the context under investigation and the original context come to light. Consequently, these differences might be utilised to estimate the effect in the new context and identify where the policy has to be altered to obtain the same effect (Williams, 2020).

Lastly, including structural causal models can increase external validity. Structural modelling is opposed to reduced form modelling, among which the CR methods are generally considered. Structural models calculate economic relations by identifying the causal mechanisms and collecting these in a mathematical model. Reduced form modelling estimates economic relations based on empirical data. As the methods of the CR are reduced form models, they lack a particular structure which is a disadvantage regarding the external validity (Deaton and Cartwright, 2018). One might interpret this structure as a framework that models multiple economic relations to try to mimic the actual economy. A researcher might extrapolate this model to another economy and tweak the variables that are different; in this way, some external validity might be achieved.

Furthermore, Deaton and Cartwright argue for creating a broad knowledge base consisting of a combination of the obtained results, the knowledge already present on the specific topic and the information from an RCT. Their study aligns with recent attempts that combine structural and reduced estimation. The idea is to use the results obtained from a reduced-form model, such as a quasi-experimental study, to calibrate parameters in a structural model. However, a superficial structural model still does not improve the external validity of the results (Angrist and Pischke, 2010). On the other hand, a holistic structural model informed by both qualitative data and unobservables might improve the external validity. This model should include contextual factors that are not directly observable but still important in determining the magnitude and particularities of the causal effect.

5 Positivism

To improve our understanding of external validity and enable ourselves to provide adequate solutions, we need to consider a philosophical theory that provides the basis for tackling the problem of external validity. Moreover, this theory or framework should be capable of facilitating the possible solutions discussed in the previous section. Considering philosophical theories is essential for tackling the problem of external validity as they provide conceptual frameworks and perspectives that help us to comprehend and assess the generalisability and relevance of scientific findings beyond specific contexts. Moreover, these theories can provide us with the intuition to improve our understanding of reality and show us on what grounds statements on scientific findings can be justified and validated.

Generally, there are two kinds of statements in economics: positive and normative. Positive statements are concerned with what is, was or will come about. These statements are descriptive statements on the economy. On the other hand, normative statements are concerned with what ought to be and require value judgements on how the economy should look. In his seminal paper, Milton Friedman argues that economics should essentially be a positivist science (Friedman, 1953). He stresses the importance of empirical evidence and scientific methods to obtain valid economic predictions. Furthermore, Friedman argues one should not judge economic methods by their ability to describe economic phenomena accurately but by their ability to predict and inform policy analysis. This positivist approach to economics originated around 1920 with the rise of the Vienna Circle. The Vienna Circle was a group of scholars with backgrounds in philosophy and the natural sciences who met regularly to discuss the formalisation of the natural sciences. A result of these meetings is the philosophical framework called logical positivism. Their goal was to create a formal demarcation principle between empirical science and meta-

physics, inferring which statements might are considered scientific or meaningless (Caldwell, 1980). They argued that all meaningful scientific statements should adhere to a principle of verifiability. According to this principle, a statement is meaningful if it can be tested against empirical evidence. However, this criterion of testability and verifiability poses many problems as it voids the meaning of theoretical entities such as radio waves or sub-atomic particles (Caldwell, 1980). Hence, this strict stance on verification created immense difficulties for formulating general laws and, according to Karl Popper, annihilated much of the natural sciences (Popper, 1959). Therefore, Popper devised his principle of falsifiability as a demarcation criterion. This principle holds that a statement becomes meaningful if it can be falsified by empirical evidence. In conclusion, the priority of the (logical) positivist approach is to obtain scientific knowledge by collecting empirical evidence.

5.1 Positivism in the Credibility Revolution

As discussed earlier, the methods from the CR aim to quantify causal relations in economics. What does the positivist view of economics entail for external validity? The methodological approach induced by positivism shares essential similarities with the approach in the CR, namely the recognition of the importance of testing statements against empirical evidence. For example, take the theory of demand and supply in the labour market, essentially tested in Card and Krueger (1993). They test this theory against the empirical evidence obtained from a natural experiment. A consequence of the theory of supply and demand theory in the labour market is that when the price of labour increases, the supply should decrease. The paper tests this statement and concludes that there is no empirical evidence to assert this statement. Hence, according to positivism, the theory is falsified as the verifiability principle does not hold. In this respect, one might approach the CR via positivism as they both emphasise the importance of verifying or falsifying scientific statements. Moreover, they agree that empirical analysis should be the basis for this verification. The quasi-experimental methods assess the observable data and try to identify and quantify causal relations just as positivism provides a framework which quantifies and objectifies scientific statements. In this sense, positivism neatly fits as a conceptual framework for the CR. In addition, a positivist approach is also an approach to searching for objective generalisable truths, which corresponds to the quest for external validity in the CR.

However, problems arise as the unobservables are not recognised as meaningful since they cannot verify them empirically. The possible solutions to the problem of external validity outlined in section 4.1 require a formal recognition of unobservable phenomena. Besides, tackling

the problem of external validity requires looking at contextual and subjective factors, such as beliefs or cultural elements. These factors are hard to measure empirically and often go beyond the specific theory of the experiment. Causal relations are embedded in social contextual circumstances, and cultural and institutional elements influence the exact causal relations. Not recognising these will not only problematise external validity but also result in a miscomprehension of reality. Later positivism does allow for unobservables if they are embedded in a theory which can be verified. This approach recognises the meaning of theoretical concepts and unobservable phenomena, albeit only concerning scientific explanation. Nevertheless, positivism mainly addresses the problem of verification and confirmation of scientific theories, whereas external validity concerns the generalisability of scientific results. Therefore, positivism is not a suitable framework to tackle the problem of external validity.

6 A realist Credibility Revolution

As discussed in the previous section, a philosophical framework should at least recognise and accept the existence of unobservable phenomena. I argue in the previous section that positivism is incapable of such a task. A philosophical theory ascribing meaning to unobservables and providing an alternative to (logical) positivism is realism. Therefore I consider realism as a possible framework. The concept of realism comprises many forms such as scientific realism, TR, empirical realism etc.[5] I choose to focus on the concept of TR described extensively by philosopher and economist Tony Lawson (TL). In his book *Economics & Reality*, Lawson argues that economists are using unsuitable methods for their subject matter and should adopt a broader approach that prioritises contextuality and is more critical. This paper builds further on Lawson's work and utilises his approach of TR to investigate the philosophical foundations of external validity. Consequently, I argue why this framework is suitable for the CR and aids in tackling the problem of external validity.

6.1 Transcendental Realism

Lawson considers TR an answer to the empirical views laid down by logical positivists. He draws heavily on Roy Bhaskar's work and his theory of critical realism.[6] To begin with, Lawson identifies a distinction between TR and empirical realism in two critical manners.[7] First, TR

[5] See the Wikipedia article for all kinds of realism: https://en.wikipedia.org/wiki/Realism

[6] See: Bhaskar (1975, 1998)

[7] Empirical realism for Lawson entails: "the claim that reality is exhausted by atomistic events and their constant conjunctions" (TL, p.20). This coincides roughly with the definitions of (logical) positivism I gave in section 5

adds to the notion of the world consisting of events and states from our experiences, the existence of underlying structures and mechanisms that exist independent of our observation (TL, p.21). The latter conduct and aid the observable events. The second difference is that the multiple levels of reality are unsynchronised. To understand the first difference, I carefully explain all concepts introduced by Lawson. A structure is the composition of a complex thing, such as systems and structured situations. Lawson argues that the world is partly composed of these complex things that aid certain events and actions in light of their structure. These complex things exist independent of their activation, i.e. they have the potential to perform their action at all times, which becomes known to us if we investigate their structure. For example, a scooter can assist us in going anywhere, independent of whether it is parked in the shed or we are currently driving it. These structured things work in a certain way, called a mechanism. A mechanism's causal power depends on having the power to act in its specific manner. Humans possess the capability to activate these powers. The activation of such a power represents the causal power of a structured thing to influence the phenomenal world. The scooter drives us to our friend when we start the scooter and accelerate. The causal power of the scooter is then the movement of us and itself.

Lawson argues that reality consists of three distinct domains: the empirical, which is grounded in our experience and impressions; the actual, which is the domain that, in addition to the empirical, contains the events and state of affairs and finally, the real domain which, besides the former domains, contains the structures, powers, mechanisms and tendencies (TL, p.21).

Lawson describes the falling of a leaf to the ground; the three domains are present here in the sense that we experience a leave falling to the ground, the leaf falls to the ground and underlying this movement are the real concepts, structures and mechanisms such as gravity (TL, p.22). These domains cannot be reduced nor identified from one another, i.e., they are unsynchronised. Hence, the possibility arises for events to be interpreted in multiple ways as the domains are not linked. In other words, the mechanisms are independent of their corresponding event. Furthermore, events are caused by multiple factors which are difficult to identify individually. Tendencies, then, are individual effects which are not directly identifiable. These tendencies are present at all times, albeit not always manifested. Consequently, tendencies are present regardless of the specific effect or context. Returning to the example of the leaf, the tendency of the leaf is the gravitational power which is present regardless of whether the leaf falls to the ground or is flying through the air in a gust of wind (TL, p.22).

According to Lawson, the scientific endeavour is not merely concerned with establishing verifiable scientific laws; the goal is to describe and shed light on the underlying structures and

mechanisms (TL, p.23). Scientific explanation is now concerned with explaining the structures and mechanisms that underlie empirical phenomena. Lawson calls this kind of reasoning reproduction, which entails a shift from the empirical to the real. Science aims to lay bare the mechanisms and structures that are the foundations of a tendency that expresses itself in the phenomenal world. Statements about regularities concern the mechanisms that unconditionally trigger a certain tendency. However, tendencies do not show themselves straightforwardly but are often conjoined with multiple tendencies. This dependence on other tendencies influences how the phenomena show itself to us. Concerning the progress of science, Lawson notes that science can only go deeper into the real phenomena left to be explained. Each explained phenomenon still covers an unexplained phenomenon underneath it (TL, p.23).

In the following chapter, Lawson argues why his TR is more suited than the positivist endeavours (TL, p.27). He establishes that one of science' aim is to search for event regularities, such as causal relations. However, in the natural sciences, these regularities are found under strict experimental controlled conditions.[8] Consequently, the experimenters themselves often bring about the causal relations observed. Nevertheless, these uncovered regularities are, at least in the natural sciences, often successfully applied outside the original experimental context. This extrapolation is essentially the problem of external validity since it concerns the generalisability of experimental results. As a result, the seldom naturally observed regularities create the problem of where our general natural laws derive meaning from (TL, p.27). According to Lawson, these general laws should not be considered event regularities but intransitive and structured scientific objects (TL, p.28). Accordingly, the results obtained through experiments and their application in other contexts require accepting a TR view where the actual events are governed by the structures, mechanisms and tendencies underlying these events. This recognises the existence of unconditional ever-present tendencies. Furthermore, Lawson understands experimental activity as controlling for possible confounding variables such that a specific mechanism is triggered. However, this experimental intervention does not uncover an empirical law; it merely sets the conditions to identify the unobservable tendencies or mechanisms considered natural laws. Consequently, these natural laws are always present but are only made known when the conditions are satisfied. Thus, generalising the results to other contexts might be viable as these laws are ever-present (TL, p.29).

Lawson advances by arguing that, in the social sciences, including economics, such regularities as found in the natural sciences are yet to be encountered or discovered. So how does Lawson argue for a TR conception of science in the social sciences? Lawson recognises the issue

[8]Lawson stresses that the field of astronomy does not adhere to this observation. Regarding the explanation of this exception, I refer the reader to (TL, p.29)

as the presence of human decision-making, implying that outcomes could have been different. He calls this the open nature of human action (TL, p.30). Subsequently, he proves the existence of a social structure that is dependent on human action and, as such, cannot be reduced to one another. This social structure only shows itself through human intervention. Regarding economics, strict event regularities are called economic laws and are defined as tendencies which seldom obtain the regularity of a natural law, i.e. economic laws break down easily as a result of contextual differences (TL, p.31-32).

The seldom appearance of law-like regularities in economics creates problems for the logical possibility of external validity of experimental methods. How could one generalise experimental results when an obtained result breaks down when extrapolated to other contexts? In order to do this, one would need to prove the existence of event regularities in the field of economics. Lawson provides conditions on a system that, if satisfied, ensure the existence of event regularities (TL, p.77). Below I show that to obtain these empirical regularities, a TR view of the methods in the CR should be accepted. First, I introduce the conditions as proposed by Lawson. Following, I argue why a TR view is required such that the methods of the CR satisfy these conditions.

6.2 Empirical regularities in economics

The first condition Lawson identifies is the extrinsic closure condition (TL, p.77-78). This condition entails that any factors or variables that might influence the dependent variable are restricted so that they do not influence the event regularity. These restrictions create a closed system where all possible influences are accounted for or uncorrelated with the specified conditions. When we translate the condition proposed by Lawson from philosophical language into econometric language, we obtain the following condition: The specified model controls for the variables that affect the outcome variable, and the variables not included in the model should be uncorrelated with the specified model. The former is considered the endogeneity problem as explained in section 3. Take the simple regression model $Y = \beta X + \epsilon$ where X is a vector containing the explanatory variable and possible control variables, β is a vector containing the treatment effect and the coefficients of the control variables, and ϵ represents the error term. With the inclusion of relevant control variables, the first point is satisfied. Regarding the second point of endogeneity, we saw that by applying the method of IV, the identification of the causal effect is still possible while accounting for endogeneity. Including controls in the method of DiD also satisfies the first point of controlling for variables that might affect the outcome variable. Furthermore, remember that DiD relies on a natural experimental setting which randomises the treatment assignment. If this assignment is random, the treatment variable is, by definition, un-

correlated with other variables not accounted for; hence, the second point is also satisfied. The same arguments apply regarding RDD as it requires the same random assignment of treatment as DiD, and control variables are included easily. Therefore, these methods of the CR satisfy the condition of extrinsic closure.

Lawson proceeds with the intrinsic closure condition, but for the sake of argument, I first discuss the aggregation condition. This condition entails that econometricians often base their analysis on micro-level data but aggregate their outcomes to a macro level (TL, p.80-81). To sustain this aggregation, some aggregational conditions should hold to combat the heterogeneity present among subjects. Individuals act differently; thus, an aggregation consists of heterogeneous individuals except if conditions apply such that a certain degree of homogeneity is present. Another solution is to ensure the compositional stability of research subjects. All three methods explained in section 3 do not address the problem of aggregation; they merely estimate effects within specific subgroups. However, in combination with a stratified sampling approach, these methods can satisfy the aggregation condition. Stratified sampling divides the population into subgroups which share similar characteristics. This split enables the researcher to generalise their results by identifying specific groups. Combining stratified sampling with the CR methods satisfies the aggregation condition as a degree of homogeneity is obtained, and integrability conditions can be met. The former entails conditions on which one might generalise the results from individuals.

Finally, Let us look at the intrinsic closure condition (TL, p.78-79). According to this condition, a researcher should take restrictions to ensure that the influence of the subjects of analysis on the outcome variable is constant. The researcher should specify the possible states of individuals to meet the requirement of an intrinsically stable structure. However, Lawson argues that for this condition to hold, the concept of reducibility must also be verified. Reducibility holds when a specific intrinsic structure of a subject can only lead to one outcome. Moreover, it must be able to express the outcome in constant terms of the intrinsic structure of a research subject (TL, p.78-79). Individuals contain a particular intrinsic structure and complexity, which, if the intrinsic closure condition holds, remains unchanged independent of any changes in the extrinsic context. The internal complexity of humans stems from their intentionality; individuals behave differently when put in other contexts. As social relations are ever-changing, the context individuals find themselves in will also change. While the description of this internal complexity might be attempted by including all sorts of individual variables, it is implausible that this captures the entire internal complexity. Thus, it does not seem that the quasi-experimental methods of the CR satisfy this condition. Such a restriction on the behaviour of the research

subjects is possible in a laboratory setting; however, in the natural settings of quasi-experimental methods, the behaviour cannot be controlled. Another option is not to control the behaviour but to try to identify how the behaviour changes in a different context. The proposed adaptions from section 4.1 try to include the influence of internal structures and mechanisms in the causal model. For example, the mechanism mapping method identifies all contextual assumptions and tries to identify how the causal mechanisms change when the contextual assumptions change. The identification of causal mechanisms requires a deep dive into each step in the causal chain, which depends on human behaviour. The researcher should consider contextual assumptions of social phenomena and the corresponding structures and mechanisms present next to the causal effect. As a result, this approach maps a specific human behaviour to a specific context, thereby achieving the reducibility principle. Alternatively, consider the speculation approach proposed by Banerjee et al. (2017). While less formal than mechanism mapping, this approach encourages researchers to speculate about the intrinsic structures of the individuals. It is unlikely that such an approach will satisfy the principle of reducibility. Nonetheless, it exerts researchers in the right direction. The verbal model proposed by Laurent (2000); Forrester (1997) is more likely to satisfy the reducibility principle. Through qualitative research, data is obtained on the essential variables and relationships characteristic of the intrinsic structure.

However, note that both the satisfaction of the intrinsic closure condition and the applicability of the methodological adaptations depend entirely on accepting a TR view. Consequently, the possibility of establishing empirical regularities in economics and the validity of the possible solutions to tackle external validity is possible only if economics embraces a TR framework. As explained, TR allows and accepts the existence of unobservable elements such as structures and mechanisms. It promotes science as uncovering the underlying mechanisms and structures. Without these underlying mechanisms and structures, it is impossible to obtain external validity for the methods in the CR. First, the proposed solutions tackle external validity by resorting to unobservable elements. The approach is to provide a holistic framework consisting of the actual, the real, and the empirical. The actual is the study of the events that happened and reporting these findings. The empirical entails collecting data and information such that analysis can take place. Moreover, the real consists of mapping individuals' contextual assumptions and behaviour and identifying causal mechanisms. To arrive at these, the researcher needs to dig as deep as possible to attempt an explanation of the phenomena underneath the apparent causal relations. Therefore, these solutions can only be fruitfully applied by accepting TR. Besides, I argue that to create the possibility of obtaining external validity, the possibility of empirical regularities in economics must be proven. The methods surrounding the CR satisfy the condition of extrinsic

closure and, in combination with stratified sampling, the aggregation condition. However, the intrinsic closure condition requires the possible methodological adaptions discussed and consequently depends on TR. Therefore, a TR perspective of econometric methodology is necessary to obtain the possibility of external validity for the methods in the CR. I must stress again that in no sense do I have the illusion to have solved the problem of external validity for the CR, nor that by accepting TR, the problem of external validity simply dissolves. I aim to create the possibility of obtaining external validity and propose the possible methodological adaptions to accomplish this for the CR.

7 Conclusion

This paper investigated the philosophical underpinnings of the concept of external validity. It argued for the necessity of accepting TR as a philosophical framework to tackle the problem of external validity in the CR. After an extensive explanation of the CR and its methods, an analysis followed of the concept of external validity. Furthermore, possible methodological solutions for the external validity problem are investigated and explained. Subsequently, I explored positivism as a viable philosophical framework to combat the problem. While positivism can provide a framework for the CRs methods, it cannot accommodate the solutions for external validity. The problem is that positivism mainly describes empirical economic phenomena, and external validity requires investigating the underlying unobservable phenomena. Therefore, TR was considered as it enables the logical possibility and facilitates methodological solutions for obtaining external validity. TR recognises the importance and existence of unobservable phenomena and can prove the existence of empirical regularities within economics. Moreover, possible methods to obtain external validity, such as speculation (Banerjee et al., 2017), verbal modelling (Forrester, 1997; Laurent, 2000) and mechanisms mapping (Williams, 2020) all include unobservable social contextual phenomena in their model to improve the generalisability of experimental results. Consequently, these methods require TR, as unobservable social and contextual phenomena are grounded in these methods. Therefore, I have shown the necessity of embracing TR to acquire external validity in the CR. Embracing TR is essential for tackling the problem of external validity as it provides a conceptual framework and perspective that helps us to comprehend and assess the generalisability and relevance of scientific findings beyond specific contexts.

Note that I do not claim to have proven that TR dissolves the problem of external validity. I have shown merely the logical possibility as well as the viability of possible solutions to depend on TR. Whether it is possible to obtain external validity in practice is subject to further

research. Methods keep improving and getting closer to the goal of external validity. The policy information will benefit if the literature concerning external validity increases and sufficient methodological improvements are made. For example, researchers should investigate the nature of the underlying unobserved phenomena that govern economic relations. When novel research sheds more light on these phenomena, it increases the understanding and practical possibilities of generalising evidence to different contexts. Finally, as noted, the existence of empirical regularities is influenced by human volition. Essentially, this discussion depends on the concept of free will. If free will were non-existent, one could argue that economics would become a natural science, as all structures would be determined by natural phenomena. Further research is necessary on the topic of human volition in economics, as it will have a significant impact on the subjects discussed in this paper.

8 Bibliography

J. D. Angrist and J.-S. Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24 (2):3–30, 2010.

A. V. Banerjee, E. Duflo, and M. Kremer. The influence of randomized controlled trials on development economics research and on development policy. *The state of Economics, the state of the world*, pages 439–498, 2016.

A. V. Banerjee, S. Chassang, and E. Snowberg. Decision theoretic approaches to experiment design and external validity. In *Handbook of Economic Field Experiments*, volume 1, pages 141–174. Elsevier, 2017.

N. Bardsley. Sociality and external validity in experimental economics. *Mind & Society*, 9(2): 119–138, 2010.

M. Bertanha and G. W. Imbens. External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 38(3):593–612, 2020.

R. Bhaskar. *A realist theory of science*. Hemel Hempstead: Harvester Press, 1975.

R. Bhaskar. Philosophy and scientific realism. *Critical realism: Essential readings*, pages 16–47, 1998.

B. Caldwell. Positivist philosophy of science and the methodology of economics. *Journal of Economic issues*, 14(1):53–76, 1980.

D. T. Campbell. Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, 54(4):297, 1957.

D. Card. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research, October 1993. URL http://www.nber.org/papers/w4483.

D. Card and A. B. Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

T. Cook and D. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, 1979.

B. De la Cuesta, N. Egami, and K. Imai. Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis*, 30(1):19–45, 2022.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

A. S. Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report, National Bureau of Economic Research, 2009.

J. W. Forrester. Industrial dynamics. *Journal of the Operational Research Society*, 48(10): 1037–1041, 1997.

M. Friedman. The methodology of positive economics. 1953.

F. Guala. The problem of external validity (or "parallelism") in experimental economics. *Social science information*, 38(4):555–573, 1999.

F. Guala. Experimental localism and external validity. *Philosophy of science*, 70(5):1195–1205, 2003.

F. Guala and L. Mittone. Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(4):495–515, 2005.

J. Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, pages 441–462, 1997.

J. J. Heckman. *Selection Bias and Self-selection*, pages 201–224. Palgrave Macmillan UK, London, 1990.

C. Heij, P. de Boer, P. Franses, T. Kloek, and H. van Dijk. *Econometric Methods with Applications in Business and Economics*. Oxford University Press, United Kingdom, 2004. ISBN 0199268010.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

G. Laurent. Improving the external validity of marketing models:: A plea for more qualitative input. *International Journal of Research in Marketing*, 17(2-3):177–182, 2000.

T. Lawson. *Economics and reality*, volume 9. Psychology Press, 1997.

E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1): 31–43, 1983. ISSN 00028282. URL http://www.jstor.org/stable/1803924.

M. L. Meldrum. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America*, 14(4):745–760, 2000. ISSN 0889-8588. doi: https://doi.org/10.1016/S0889-8588(05)70309-9. URL https://www.sciencedirect.com/science/article/pii/S0889858805703099.

K. Popper. *The logic of scientific discovery*. Routledge, 1959.

J. Reiss. *Philosophy of economics: A contemporary introduction*. Routledge, 2013.

M. R. Rosenzweig and K. I. Wolpin. Natural "natural experiments" in economics. *Journal of Economic Literature*, 38(4):827–874, 2000.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

J. Snow. On the mode of communication of cholera. *Edinburgh medical journal*, 1(7):668, 1856.

S. Urzua, J. J. Heckman, et al. Comparing iv with structural models: What simple iv can and cannot identify. Technical report, 2009.

M. J. Williams. External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer*, 35(2):158–191, 2020.

C. Wing and R. A. Bello-Gomez. Regression discontinuity and beyond: Options for studying external validity in an internally valid design. *American Journal of Evaluation*, 39(1):91–108, 2018.