Methods for impact evaluation

(PEN Deliverable 3.1)

Dr Sara Capacci[1]

Professor Mario Mazzocchi[1]

Dr Laurent Muller[2]

Dr Beatrice Biondi[1]

On behalf of the PEN consortium

[1] Alma Mater Studiorum – Università di Bologna, Department of Statistical Sciences
Via Belle Arti 41, 40126 Bologna, Italy, http://www.stat.unibo.it/en

[2] Grenoble Applied Economics Laboratory, French National Institute of Agricultural Research, 1241, rue des residences, 38400 Saint Martin d'Hères, France, https://gael.univ-grenoble-alpes.fr

Capacci S, Mazzocchi M, Muller L, Biondi B. Methods for Impact Evaluation - Deliverable 3.1. Bremen, Germany: JPI HDHL Policy Evaluation Network, 2022: 4–12

Methods for impact evaluation

1. ABSTRACT

Deliverable 3.1 “Methods for Impact Evaluation” reports the activities and research findings of sub-tasks 3.1.2 and 3.1.3 of the PEN project.

The focus of this Deliverable is on methods for the quantification of policy impacts in terms of efficacy and effectiveness. It responds to two key needs: (a) identifying the strengths, weaknesses and the main conditions required for a rigorous ex-ante evaluation of policy impacts through lab-field-natural **experiments**, especially in relation to capturing behavioural phenomena; and (b) reviewing the applicability and limitations of **quasi-experimental econometric methods** for the ex-post evaluation of policy impacts based on observational data not collected for evaluation purposes, and case studies.

The Deliverable provides an overview of the use of:

- experiments for policy evaluation, with a discussion of the methodological challenges. Results of the comparison between the framed-field experiment and the natural-field experiment on nutritional labelling in France, that were presented during the PEN-WP3 workshop in September are presented in a detailed case-study analysis.
- quasi-experimental methods for policy evaluation, through a detailed report (Appendix 1) on the methodological challenges for the various approach. The report includes a discussion of the two case studies (Catalunya sugar-sweetened beverage tax, and Cycling May campaign in Gdansk) accompanied with Stata codes and data to replicate the results. A Stata do file (Syntax.do) and two Stata dataset (CPI.dta, Price.dta) were part of this Deliverable, requests to see this material can be sent to jpi-pen@leibniz-bips.de.

2. THE USE OF EXPERIMENTS FOR POLICY EVALUATION

Experiments are increasingly used to help public and private decision-making on societal issues. In particular, they are a common methodology used in consumer research to better understand and predict the behaviour of individuals in the marketplace. Experiments allow for the *ex-ante* evaluation of an action; i.e. Experiments make it possible to analyse economic phenomena for which there is no (or not yet) data or for which data are difficult to observe or exploit. This is particularly useful when, for instance, policy makers are looking for arguments to support the relevance of a policy they are considering launching in the future. We will later see the example of the French Health Ministry, which conducted in 2016 experiments to determine the best format for nutrition labelling, resulting in the Nutri-Score on the front of food packaging in France and some other European countries.

In the mix of evaluation methods, experiments serve as the middle ground between theoretical simulations and real-world observations (see Table 1). On the one hand, simulation-type studies have a high degree of internal structure and validity but may lack behavioural realism. On the other hand, observational studies examine actual behaviour in real-life situations, but they often lack the level of control needed to conclusively identify causal effects. In this regard, experiments offer an interesting alternative. First, the results of the experiments are derived from actual behaviours rather than from deductive-hypothesis inferences in simulation-type studies or from declarative statements in survey or focus group studies. Second, experiments differ from observational approaches in that they are inherently designed to compare outcomes between groups (treatment groups *vs.* control group), thus allowing the researcher to identify a causal relationship between an intervention and a respondent's reaction.

Nevertheless, the tension between external and internal validity still exists even across the spectrum of experimental methods. “Where internal validity often requires abstraction and simplification to make the research more tractable, these concessions are made at the cost of decreasing external validity” (Schram 2005). Following a long tradition of deductive reasoning and modelling in economics, internal validity is quite rightly the utmost priority for most economists. This is how laboratory are carried out in a controlled environment with rigorous design protocols and relatively small samples. The main objective here is to provide results that can be replicated to support or reject existing or emerging theories. Although theory should always guide economists, external validity is of major importance when the objective is to advise policy makers (“whispering to the ears of princes”, Roth 1995). In order to improve prediction, the experiments must then incorporate key characteristics that are specific to the market under study. Thus, field experiments are carried out in a less controlled environment but with higher ecological validity.

Hence, the methodological challenge is to devise experiments that better mimic real-life situations without compromising the control of the explanatory variables. In the following, we first describe experiments in economics and how they differ from experiments is psychology. Second, we discuss the internal and external validity of experiments. Finally, we look at the case study of experiments used to evaluate the impact of nutrition labelling systems on the front of pack of food.

		EXPERIMENTS					
Type of study	NUMERICAL SIMULATIONS	**Conventional laboratory experiment**	**Artefactual field experiment**	**Framed-field experiment**	**Natural-field experiment**	NATURAL DATA	SURVEYS and FOCUS GROUPS
Intervention examined	Artificial exogenous shock					Naturally-occuring events	*“Let us pretend”*
Sample	Fictious character (*e.g.* the *Homo Oeconomicus*)	Students	Target population				
Nature of the task	Abstracted situations			Contextualized situations			
Setting	*In Vitro*				*In Vivo*		
Nature of the answer	Deductive-hypotheses inference	Behaviours					Statements

***Table 1**. Description of methods used to evaluate policy impacts*

a. Experiments in economics

An experiment in economics consists of reconstituting a simplified economic situation in a controlled environment. The experimenter chooses the variables of interest as well as other variables, either explanatory or control variables, in order to determine the impact *ceteris paribus*[1] of these on the individual or collective behaviour of the participants. Such a control makes it possible to identify the causal effect of these variables.

Experiments serve one or more of the three purposes defined by Davis and Holt (1993) and Roth (1988): (i) to test theory in order to identify mechanisms not predicted by existing models; (ii) to produce facts in order to

[1] This consists of varying the variable under study when all other variables are held constant.

identify previously unknown behavioural regularities that can be incorporated into theoretical models; (iii) to aid decision making by testing the implications of implementing a new mechanism. The first two objectives are interdependent and mutually supportive. The third objective is to assess the economic consequences of an organisational or institutional change. It is thus possible to assess the impact of the implementation of various public policies, different forms of market regulation or different organisational systems within the company.

Most of the experiments found in consumer-focused journals follow the experimental norms used in psychology. While experimental economics draws heavily on experimental psychology, conventional laboratory experiments in economics (Table 1) are distinguished by the inclusion of four basic practices: abstraction, stationary replication, consequential incentives and the absence of deception (Huff 2014).

(i) Abstraction

Abstraction means using general and neutral terminology in the instructions and in the stimuli used during the experiment (Davis and Holt 1993). The idea is to employ the most parsimonious procedure possible in order to more easily isolate the variables of interest. This is particularly useful for testing theories. Indeed, abstract terminology simplifies the experimental context and thus makes the conclusions of the theory easier to observe and test. The rationale is that if a theory does not hold in its simplest form, it is unlikely to hold in the more complex real world.

Another benefit of abstraction is that it helps researchers to replicate and compare results. It allows experimentalists across the world to design very similar protocols. By employing the same terminology and measuring the same dependent variables, it contributes to the generalizability and the robustness of the findings (Camerer 2003).

(ii) Stationary Replication

Historically, economists have sought to predict stable or equilibrium behaviour (Camerer 1997). In other words, they prefer to know what decision a person will typically make after being confronted with the decision several times rather than what that person will decide when confronted with the decision the first or second time. Therefore, experimental economists typically conduct repeated trials. In a single session, participants make several decisions in succession under strictly identical conditions. It allows the study on how behaviours converge toward the prediction of a theory.

(iii) Consequential Incentives

For most economists, incentives are necessary to reveal the truth (Kagel and Roth 1995). For Plott (1982), laboratory and field experiments are useful in economics because they reflect a real decision process with real people performing defined tasks and whose decisions and efforts on those tasks have real consequences. A proper incentive design follows the principles of *salience* and *dominance* (Smith 1982). Salience requires that payoffs are dependent on the choices available. Dominance ensures that payoffs are high enough that participants have an incentive to do their best to complete the experimental task. When salience and dominance are adequate, an experiment is said to be incentive compatible: Participants get the best results for themselves simply by acting on their true preferences.

Consequential incentives essentially prevent the now well documented hypothetical bias (see for example Loomis, 2011). Hypothetical bias is defined as the gap between the response provided by a participant in a hypothetical experiment and what he or she would have indicated in a real or incentivised situation. There are many reasons for the existence of hypothetical bias. One of them is social desirability or also called demand artefacts. Participants are deliberately or even unconsciously tempted to answer in the most socially acceptable way. Participants may also fail to reveal their true preferences in hypothetical situations because they find it difficult to project themselves onto the task proposed by the experimenter. Finally, they may answer strategically in the belief that their answers will have real consequences, for example on questions relating to the pricing of a product or service, the implementation of a public policy, or the marketing of a new good (Lusk, McLaughlin, and Jaeger 2007).

(iv) The absence of Deception

Experimenters have long advocated against the use of deception in their protocols. Deception is actually banned in the scientific journal *Experimental Economics* and in a large number of renowned economic journal. The reasons are twofold: incentive control on the one hand and negative externalities for the profession on the other (Cooper 2014). First, the credibility of the instructions must not be questioned by the subjects. If they believe that they will be paid on the basis of criteria other than those stated in the instructions, the experimenter loses control of the incentives and the hypotheses formulated are no longer valid. Secondly, subjects who have been deceived in a previous experiment will doubt the instructions when participating in future experiments, even if those experiments do not use deception.

Experiments in consumer research are not typical economic experiments. First, consumer researchers traditionally seek a realistic experimental context rather than abstraction. This might be explained by the fact that external validity is often favoured over internal validity. Second, they very rarely use stationary replication, even though consumption activities are often repeated several times in a relatively short period of time. Third, although the use of consequential incentives in consumer studies is more common than the use of abstraction and stationary replication, it is surprising that the vast majority of consumer studies do not use them. Appropriate incentives offer both greater traction (internal validity) and greater realism (external validity) and, as such, are fully compatible with consumer research. Finally, deception is still used in consumer research, mainly in the form of cover stories, but its use is declining significantly.

Overall, experiments in consumer research are increasingly adopting the best practices of conventional laboratory experiments. The most difficult methodological challenge is certainly to strengthen the internal validity of the results without weakening the external validity.

b. Internal and external validity

Experimental economics is based on the principle of control, both of the variables and of the environment. This control is essential to study the causal impact of a variable on individual behaviour and thus to have strong internal validity. Conventional laboratory experiments allow a high degree of control. However, this control is often at the expense of a strong generalizability of the results (i.e. strong external validity). Indeed, by keeping the experimental setting free of potential noise, the laboratory can produce artefactual behaviour that is not replicated in ecological contexts, thereby undermining external validity. Conversely, the closer the experimental setting is to the real world, the more difficult it is to identify and isolate explanatory variables.

To define these notions of validity more precisely, we take up the summary presentation of Falk and Heckman (2009). Suppose that a researcher wishes to determine the effect of a variable denoted X_1 on a variable denoted Y, knowing that there is a function f that links the two variables as follows: $Y = f(X_1, X_2, \dots, X_n)$. The researcher thus studies the effect of variations in the variable X_1 on the variable Y given a given level of the other variables $\breve{X} = (\breve{X}_2, \dots, \breve{X}_n)$. We can then define internal validity as the researcher's level of control over the $\breve{X}$ variables when determining causality so that it is the effect of variable X_1 on the variable denoted Y that he is measuring. And we can define external validity as the degree to which causality is maintained between these two variables both for one level $\breve{X}$ of the other variables and for another level of these variables, e.g. $\tilde{X} = (\tilde{X}_2, \dots, \tilde{X}_n)$.

External validity is relatively more important for experiments searching for empirical regularities than for theory-testing experiments. This is particularly the case for policy evaluation experiments: A policy maker will want to know whether the results obtained in the laboratory will be sufficiently predictive. Formally, any causal relationship is externally valid if the causal relationship observed in the lab is also valid outside the lab. According to Guala (2002), this question of parallelism between the real world and the laboratory is thus one of the most important methodological issues in experimental economics. Several recent works try to define more precisely the conditions of this external validity and question the relevance of the generalisation of experimental results to the 'real world' (Levitt and List 2007; Kessler and Vesterlund 2015).

An experiment lacks of external validity when behaviours observed in the laboratory do not replicate in the field. External invalidity occurs when the laboratory setting (i) does not replicate all relevant real-world stimuli or (ii) generates distorted behaviour.

(i) Field experiments

External validity is first addressed by integrating the essential features of the case studied. These features relate to the sample, the task and the environment of the study. Conventional laboratory experiments usually invite randomly selected students to perform an abstract task in a classroom. Harrison and List (2004) distinguish three types of so-called field experiments (see table 1). They all differ, to varying degrees, from conventional laboratory experiments in that they successively and cumulatively address issues of representativeness, framing and ecology. By incrementally importing the target population, then the task and finally the environment into the experimental design, the observed behaviours are more and more likely to replicate those that occur naturally.

- *Representativeness*. The first step in improving the parallelism between the laboratory and the outside world is to develop a sample of participants that is representative of the population being studied. Experiments where participants are not students but the target population of the study (farmers, consumers, business leaders, etc.) are called **Artefactual-Field experiments.**
- *Framing*. The second step consists in contextualising the experimental setting by deriving decisions, products, information, etc. from the real world. Contextualized experiments with non-student participants are called **Framed-Field experiments**. Artefactual-field experiments and framed-field experiments are still carried in laboratories (*in vitro*).
- *Ecology*. The final step is to leave the laboratory walls to observe participants behaving in their natural environment (*in vivo*) in so-called **Natural-Field experiments.** Natural-Field experiments transpose the controlled procedures of conventional experiments outside the laboratory. They are also known as **lab-in-the-field studies**.

Field experiments remain experiments in the sense that, as in a traditional experimental design, the independent variable is manipulated by the experimenter. And as in conventional laboratory experiments, the dependent variables are expected to be measured under strict conditions *ceteris paribus*. It is indeed a virtue of experimental design that any potential impact of other variables (extraneous factors) is equalised between treatments. Nevertheless, the more realistic are the sample, task and environment, the more difficult it is to control for all variables. To the extent that uncontrolled extraneous variables may be impacting on the dependent variables, the causal relationship between the dependent variables and the manipulated independent variable cannot be established, thus undermining internal validity. To ensure sufficient control over confounding factors, field experiments must follow standardised procedures to warrant sound design and analysis.

Provided they are well designed, properly conducted, and enrol enough participants, **Randomized Controlled Trials** may ensure such control. Randomized controlled trials consist in randomly allocating participants among compared treatment groups and control group. The random assignment of participants to treatments reduces selection bias and assignment bias. The presence of a control group allows the effect of the manipulated independent variable to be isolated. Randomised controlled trials have become the gold standard for interventional studies. While **quasi-experimental methods** also use a control group, they differ from field experiments in that they do not use randomization. Therefore, quasi-experiments are subject to concerns regarding internal validity, because the treatment and control groups may not be comparable at baseline. Nevertheless, they may be the only option when field experiments are not feasible.

(ii) Artificiality

Artificiality of the setting is a major obstacle to the external validity (Bardsley 2005, Schram 2005). An experiment has the danger of creating its own world. The artificial nature of the laboratory situation may lead to different choice behaviours from a real context (Harrison Harstad and Rutström 2004).

First, consciously participating in an experiment may condition responses in a way that affects the causal relationship observed (Starmer 1999). Participants in experiments may change their behaviours due to cues about what constitutes appropriate behaviour. Such demand effects are about the vertical relationship between the experimenter (the expert) and the participant: Participants have to produce what is *demanded* by the experimenter. Demand effects can either be social or purely cognitive (Zizzo 2010). Participants may be sensitive to the fact that an experimenter is monitoring their behaviour (Hawthorne effect). For instance, participants may be inclined to send messages to the experimenter in order to influence the outcome of the study (strategic bias) or behave in a way that they think is expected by the experimenter (desirability bias). Also, participants may not respond to the experimenter instructions as demanded. Such misperceptions of the task can occur when participants simply do not understand the task at hand (especially when tasks are very abstract) or when they have false beliefs about what the experimenter actually wants to test. Compatible incentives, clear task construction and absence of deception are the experimenters' best weapons to ensure that the observed behaviours reflect as much as possible the participants' true preferences.

Secondly, participants in experiments are not in the same frame of mind as they usually are in everyday situations. They are generally more focused, pay more attention to external cues and are more reflective. They usually have time to make a decision and make the necessary efforts. In everyday life, they are more likely to make rash decisions and rely on heuristics. According to Kahneman (2011)'s dual cognitive system, experiments are conducive to the deliberate system 2 at the expense of the intuitive system 1.

c. Framed-field experiment vs. Natural-field experiment (Case study 1)

In 2016, two almost identical field and laboratory experiments took place in France to examine the nutritional impact of different front-of-pack labelling schemes, providing a unique opportunity to undertake a methodological comparison.

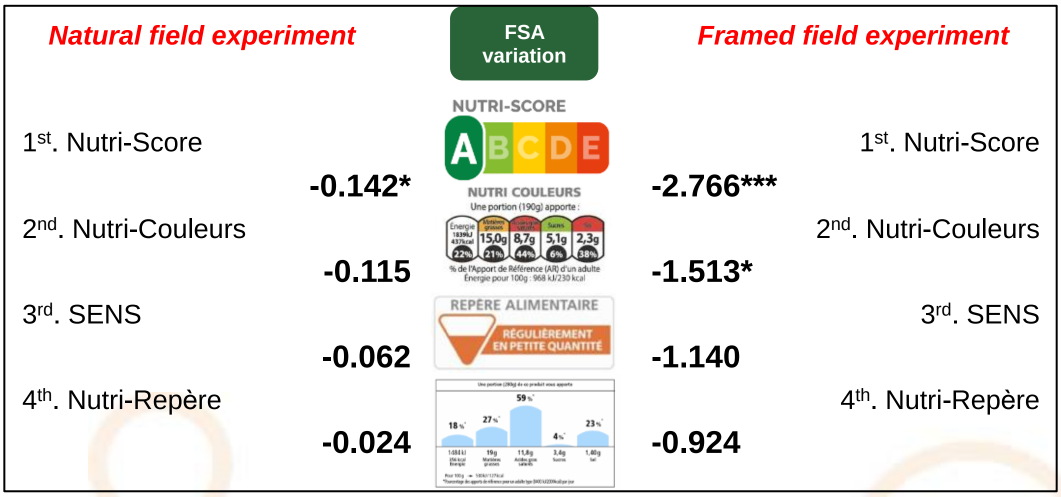
In response to the rising health costs generated by the obesity epidemic, the political response has been to pass a bill to introduce a harmonised system that will be most effective in changing consumer purchasing behaviour. The French health modernisation law of 26 January 2016 calls for a nutrition labelling system based on the nutritional composition of products. Former French Health Minister Marisol Touraine advocated the Nutri-Score, a simplified labelling format that classifies foods from A, green and healthy, to E, red and unhealthy. Her proposal triggered a heated debate among stakeholders, who questioned its effectiveness and the resulting stigma that such a label might carry. To settle the matter, the French authorities brought together all food stakeholders and launched a competition: each stakeholder was invited to propose a labelling format which will then be tested during a trial period to see which one is the most efficient in encouraging consumers healthier food choices. Four formats joined the contest: the Reference Intakes, SENS, respectively endorsed by the food industries and the retailers, the Multiple Traffics Lights and the Nutri-Score. A large natural field experiment was therefore carried out in 60 supermarkets to see which format was best for changing food purchases towards healthier diets. Given the heat of the debate[2], the Ministry of Health needed robust results that would be difficult to dispute. It therefore decided to complement the natural field experiment with a laboratory framed-field experiment. The two studies are respectively detailed in (Dubois et al. 2020; Crosetto et al. 2020)

(i) Description of the two studies

Both studies used the same experimental designs by observing purchasing behaviour before and after the implementation of a labelling scheme (difference-in-difference approach) and using the same outcome measures (FSA score normalised by 100 kcal, Rayner et al., 2009). The natural field experiment included 60 supermarkets in 4 French regions with 10 shops per system and 20 shops for control. The study lasted 10 weeks from 26 September to 4 December 2016. Consumers were informed about the local intervention in each treatment supermarket by leaflets and totems. In the labelling phase, 1266 products from four departments (Fresh prepared products, canned prepared products, pastries, industrial breads) were labelled with stickers. The coverage of the logos was between 45% and 75%, mainly of retailer branded products. The food purchases of 171,827 loyalty card holders were recorded. On the other hand, the framed field experiment took place on the experimental platform of the Grenoble Polytechnic Institute. The study included 51 sessions of 1h30 each from 21 November to 2 December 2016. 832 participants were invited to shop for their household over two days. They made their choice from a paper catalogue of 290 products. They had to perform this task twice, with and without the presence of a labelling system (All 290 products were then labelled). At the end of the session, they actually bought a quarter of their food basket.

[2] In addition to the opponents of front-of-pack labelling, there were claims (mainly from supporters of labelling) that the tests would be biased, following an investigation by Le Monde that revealed conflicts of interest.

(ii) Results



Natural field experiment		FSA variation		Framed field experiment
1st. Nutri-Score	-0.142*	NUTRI-SCORE	-2.766***	1st. Nutri-Score
2nd. Nutri-Couleurs	-0.115	NUTRI COULEURS	-1.513*	2nd. Nutri-Couleurs
3rd. SENS	-0.062	REPÈRE ALIMENTAIRE	-1.140	3rd. SENS
4th. Nutri-Repère	-0.024		-0.924	4th. Nutri-Repère

A B C D E
Une portion (190g) apporte :
Energie 1839kJ 437kcal 22%
Matières grasses 15,0g 21%
Acides gras saturés 8,7g 44%
Sucres 5,1g 6%
Sel 2,3g 38%
% de l'Apport de Référence (AR) d'un adulte
Énergie pour 100g : 968 kJ/230 kcal
RÉGULIÈREMENT EN PETITE QUANTITÉ
18 %
27 %
59 %
4 %
23 %

Figure 1. Variation the FSA score in the natural field experiment and in the framed field experiment according to labelling systems.

Both studies resulted in the same ranking, with Nutri-Score being the most effective label, i.e. the one that generates the largest decrease in FSA score. However, the size of the effects of nutrition labels was on average 17 times smaller in the natural field experiment than in the framed field experiment.

(iii) Discussion

When considering the potential causes of differences between laboratory and field studies, the literature generally points to the following five usual suspects:

- The Hawthorne effect, i.e. the nature and extent of scrutiny of a person's actions by others.
- The stakes of the decisions
- The characteristics of the sample
- The laboratory context, which may differ from the ecological context
- The nature of the laboratory task, which may not perfectly replicate decisions made in the field.

The first three arguments can be quickly dismissed. Both samples were aware that they were being scrutinised; both studies involved small stakes; the participants in both studies had similar key characteristics. Consumer attention, however, was very different. The framed field experiment examined two consecutive purchase decisions, whereas the natural field experiment examined multiple purchase decisions over several weeks. In addition, the logos were more visible in the catalogues of the laboratory study than on the shelves of the field study. Control generates saliency.

Laboratory control comes at a cost. Due to the increased consumer attention, the laboratory clearly overestimates the impact of the intervention. This contradicts Herbst and Mas (2015) who found no quantitative difference. Is this difference in effect size significant? No, if the objective is to choose the "best" option. The magnifying glass effect allows the laboratory to better discriminate the impact of competing labelling schemes. Yes, if the objective is a cost-benefit analysis. Effect size is important when simulated results are used to assess future implications for society (e.g. in epidemiology).

3. QUASI-EXPERIMENTAL METHODS FOR EX-POST EVALUATIONS ON OBSERVATIONAL DATA

Ex-post policy evaluations occur *after* the policy has been implemented and are based on the direct observation of the attribute/s of interest (i.e. no simulation is needed). Ex-post evaluations might occur under either an *experimental setting* or an *observational setting*, depending on the way the subjects are selected to be exposed to the policy (the data generating process). While an experimental setting can guarantee randomization into the policy (the data are generated by a random experiment), in an observational setting the exposure to the policy is beyond the control of the researcher who just observes the process as it takes place (the data are not generated by a randomized experiment). This latter scenario will be extensively explored in the present document.

Ex-post impact evaluation consists in assessing the change in the attribute of interest (the *outcome variable*) *caused* by the policy (the *treatment* or *intervention*). The causal effect goes beyond the simple correlation between the exposure to the policy (the treatment status) and the outcome variable and consists in the difference between the outcome observed after the policy and what would have been observed without the policy. The latter represents the “counterfactual” outcome, a theoretical quantity not observable by nature. The counterfactual is the core of the so-called fundamental problem of evaluation (Heckman et al., 1999) which makes the treatment effect not observable itself in non-experimental settings. *Quasi-experimental methods* consist in econometric and statistical techniques used to *estimate* the counterfactual scenario when subjects are not randomly selected to the treatment. They are based on statistical strategies able to reproduce as closely as possible the desirable condition of randomized experiments where two groups are identical with respect to any factor except the probability of being exposed to the policy.

While within experimental settings, randomization makes the control group a reasonable approximation of the counterfactual scenario, in observational settings, exposed and non-exposed subjects might systematically differ according to some factors, which in turn affect the outcome even in absence of the policy (the selection bias). This implies that the outcomes of the two groups would have been different even with no policy and the observed outcome differences after the policy implementation cannot be attributed exclusively to it. The magnitude of the selection bias depends on the selection process, which assigns some individuals of the population to the policy and not others. The *potential outcome framework* - variously attributed to Fisher (Fisher, 1935), Neyman (Neyman and Iwaszkiewicz, 1935), Roy (Roy, 1951), Quandt (Quandt, 1972), or Rubin (Rubin, 1974) - offers a straightforward formalization which reveals very helpful for depicting the main issues arising in policy evaluation due to the fundamental problem of evaluation and generally summarized in the selection bias problem.

Each quasi-experimental method builds on a specific identification strategy, which directly follows from the policy design and the data availability and relies on specific assumptions. So that the same methodological approach might be useful in some practical setting and may fail in others.

A structured report of quasi-experimental methods is provided in Appendix 1. Identification strategies, assumptions and data requirements of the most prominent quasi-experimental methods are illustrated and empirical discussion is provided with reference to two case studies: the soda tax implemented in Catalonia, Spain in May 2017 and the Cycling May campaign in Gdańsk, Poland.

The Catalonia Soda Tax (Case study 1)

Catalonia is one of the 19 Autonomous Communities in Spain and has high levels of fiscal autonomy. The Catalan Parliament approved the *Impuesto sobre Bebidas Azucaradas Envasadas* (IBAE) in March 2017, which came into force in May 2017 (Law 5/2017). The declared aim of the tax is to levy Sugar Sweetened Beverages (SSBs) to reduce sugar consumption and the associated harmful effects on health, responding to the WHO recommendations. According to this law, the following beverages are taxable: sodas and soft drinks, industrial fruit juices or fruit nectars, sports drinks, drinks with coffee and tea, energetic drinks, sweetened milks, shakes and juices containing milk, plant-based beverages, flavouring waters.

By law, retailers must fully pass-through the price increase due to the tax on to the final consumer. This should imply a 100% pass-through to the final prices borne by consumers.

The tax rate depends on the sugar proportion of the drink: for drinks with 5 to 8 grams of sugar per 100 millilitres, the tax amounts to 0.08 euros per litre, for drinks with more than 8 grams of sugar per 100 millilitres, the tax amounts to 0.12 euros per litre, drinks with less than 5 grams of sugar per 100 millilitres are exempt.

No similar taxes have been enforced in the rest of the country.

Available data: Data from four waves (2016-2019) of the National Spanish Household Budget Survey (HBS) provided by the Spanish National Statistics Office (INE) are available. The data include outcome observations before and after the tax entry into force for both the treated and the control regions. Regional Consumer Price Indices (CPIs) provided by INE are also used. According to the available data, the outcome variables can be measured as follows:

- Price of taxed drinks are measured
  - as Regional Monthly CPIs (non-alcoholic drinks)
  - as average price paid per region and month, computed by averaging the price paid by each household (calculated as the ratio between expenditure and purchased quantity) across regions and months.
- Purchase of taxed drinks measured as per capita annual purchased quantities.

Cycling May (Case study 2)

Cycling May is the name of a Polish public initiative aimed at increasing usage of bicycle, primarily directed to children. The campaign disseminates the idea that cycling is a fun and healthy mode of transport to reach school or kindergarten. The campaign was first enacted in May 2014 in the city of Gdańsk. In the following years, several Polish cities and municipalities adopted the initiative, reaching 47 involved municipalities in 2019.

1 Cycling May primary aim is to promote a healthy lifestyle and enhance active transport among pupils in schools and kindergartens, their parents, and teachers. The initiative is based on a competition – adopting a “gamification” approach – and includes elements of fun: each child and teacher receives a sticker every time that she cycles to school and there are final prizes for more active children, classes and schools. Each child has her own travel diary to keep track of bicycle trips, a class poster allows to record cumulative trips. Awards come in the form of bicycle gadgets for children, organised class excursions, or financial support for amelioration of schools’ cycling facilities. Aside from the competition, other activities carried out at school teach good and healthy transport habits. Interested primary schools and kindergartens must apply to participate to the initiative.

Available data: daily bicycle recording from bicycle counters located in a treated Polish city – Gdansk – and a control Polish city - Lodz - are available over a three-years period. The city of Gdansk is the capital of the Pomeranian province, a coastal province in northern Poland, where Cycling May was active. While Lodz is the capital of the homonym province, situated in central Poland, where Cycling May was not implemented. Number of daily bicycle recording at the counter level is available.

The report is structured as follows. First, the *potential outcome framework* is briefly outlined to provide the basis for a formal discussion of the selection bias problem and the general conditions for identification of the policy effect. Then a structured discussion of quasi-experimental methods follows, where each evaluation method is explored in relation to the corresponding identification strategy and selection scenario.

APPENDIX

Report on quasi-experimental methods

Abstract/executive summary

Ex-post evaluation methods are *quasi-experimental* in that they rely on observational data (where the exposure to the policy is beyond the control of the researcher) and use statistical strategies to reproduce the desirable conditions of a randomized experiment to estimate a *counterfactual* scenario (what would have happened in the absence of the policy) and ultimately the *causal effect* of the policy.

Alternative statistical methods can be used to estimate the causal effect of the policy. Among them: the difference-in-difference approach, synthetic control methods, regression discontinuity designs, model-based counterfactual estimations, instrumental variables approaches, propensity score matching. Each of them relies on specific strategies to identify the policy causal effect.

In the quasi-experimental approach, the identification strategy is far from being a mere theoretical question and it follows directly from the policy design (which determine the details of the exposure to the policy) and the data availability. So that the same methodological approach might be useful in some practical setting and may fail in others.

Identification strategy, assumption and data requirements of the most prominent quasi-experimental methods will be explored and empirical discussion will be provided with reference to two case studies.

1. Examples and applications to Case Studies

The present report consists in a structured not-too-technical discussion of quasi-experimental methods supplemented with examples and references. Applications of different methods and approaches to two Case Studies are shown, i.e. the Catalonia soda tax (Case Study 1) and the Cycling May Campaign in Poland (Case Study 2).

With regard to Case Study 2: a structured discussion of the campaign and of the methods used for quasi-experimental evaluation are provided.

With regard to applications to Case Study 1, datasets and Stata SE syntax to replicate the analyses are attached to this document. Data for the evaluation of the Catalonia soda tax come from four waves (2016-2019) of the National Spanish Household Budget Survey (HBS) and the Regional Consumer Price Indices (CPIs) provided by the Spanish National Statistics Office (INE). Using the available data and different quasi-experimental approaches we found the tax has significantly affected prices (causing a decrease in the price of taxed dinks, see details in the following sections) but it has not significantly affected purchased quantities of taxed drinks. For this reason, we decided to report examples referring to prices.

1.1 Outcomes and datasets used for applications to Case Study 1

The following outcomes are considered for impact evaluations:

- Monthly Regional Consumer Price Indices (CPIs) for non-alcoholic drinks: no further detail in terms of type of drink (e.g. soft drink, energy drink, etc.) was available at the monthly/regional level. The index refers to all non-alcoholic beverages (all taxed and non-taxed drinks).

- Price of drinks measured as average price paid per region and month. These are computed using household data from the HBS, by averaging the price paid by each household (i.e. the ratio between expenditure and purchased quantity) across regions and months. Higher detail in terms of items' aggregation is available (see below)

- Price paid for drinks (unit values): calculated as the ratio between household expenditure and household purchased quantity using household data from the HBS.

CPI.dta dataset: it contains Monthly Regional CPIs for non-alcoholic drinks recorded across 19 autonomous regions on a 48-months-period (from January 2016 to December 2019). It also includes National CPIs for non-alcoholic drinks. It consists in 960 observations. Regional/monthly demographic characteristics are included. The latter have been calculated using HBS data by averaging household demographic characteristics across regions and months.

PRICE.dta dataset: it contains average monthly/regional prices computed by averaging household prices paid per drink across regions and months. This procedure relies on the assumption that households in the same location and time period face the same price (Braha et al., 2017; Rahkovsky & Gregory, 2013). Similar procedures can be found in Colchero et al., 2015 and Beatty, 2008. Regional/monthly demographic characteristics are included as above. Original household level information (price paid and demographics) come from the HBS. The HBS data show higher detail in item aggregations with respect to CPI data. The following scheme summarizes the available item disaggregation:

- Non-alcoholic drinks:
    - Soft drinks
    - Energy drinks
    - Sport drinks
    - Juices
    - Water

The HBS does not provide any information regarding the beverages' sugar content. Thus, it is not possible to detect taxed (and untaxed) drinks with precision. The tax does not apply to artificially sweetened beverages (diet drinks) and 100% fruit juices. Moreover, different tax rates apply to beverages with different sugar proportions (for drinks with 5 to 8 grams of sugar per 100 ml, the tax amounts to 0.08 euros per litre, for drinks with more than 8 grams of sugar per 100 ml, the tax amounts to 0.12 euros per litre, drinks with less than 5 grams of sugar per 100 ml are exempt). However, the greater detail allows to focus to specific drink categories.

2. Ex-post policy evaluation in observational settings

2.1 The potential outcome framework

According to the *potential outcome framework*, being Yᵢ the outcome variable of the *i*-th subject, every *i*-th subject has potentially two outcomes Y_i^1 and Y_i^0 having received or not the treatment. Depending on her treatment status ($D_i = 1$ if the subject received the treatment and $D_i = 0$ if she did not) one of the two potential outcomes is observable and the other is hypothetical (counterfactual).

	Factual Outcome	Counterfactual Outcome
Exposed to the policy ($D = 1$)	Y^1	Y^0
Non-exposed to the policy ($D = 0$)	Y^0	Y^1

The observable outcome of individual *i* is thus:

$$Y_i = Y_i^0 + D_i(Y_i^1 - Y_i^0) \qquad 1$$

and the causal effect of the treatment for the *i*-th subject is the (unobservable) difference $Y_i^1 - Y_i^0$

It represents the outcome change that is totally attributable to switching from state $D = 0$ (no treatment) to $D = 1$ (treatment). While it is logically defined for all members of the population of interest (irrespectively from their actual exposure to the policy) and free to vary across individuals (some population members benefit more from interventions, some other benefit less, some others might even be damaged by the intervention), though it is not observable (*Fundamental Problem of Evaluation*).

The same is true for the causal effect of the treatment at the population level. The average treatment effect computed over the treated population (ATT) is defined as follows:

$$ATT = E(Y^1 - Y^0|D=1) = E(Y^1|D=1) - E(Y^0|D=1) \quad (2)$$

and is not observable (due to its second term, $E(Y^0|D=1)$, which is counterfactual). The ATT measures how much individuals exposed to the policy benefited on average from it and it is usually of major interest for policy makers.

If the (unobservable) counterfactual average outcome for the treated, $E(Y^0|D=1)$, is replaced with the (observable) factual outcome of the not treated, $E(Y^0|D=0)$, the resulting difference returns a biased estimate of ATT:

$$E(Y^1|D=1) - E(Y^0|D=0) = ATT + E(Y^0|D=1) - E(Y^0|D=0) \quad (3)$$

where $E(Y^0|D=1) - E(Y^0|D=0)$ represent the selection bias and summarizes the outcome difference between participants and non-participants if the policy was not implemented. It captures outcome differences that cannot be attributed to the policy and represents a bias to the identification of the ATT using only observable quantities.

2.2 Selection bias, randomization, and the selection process

The magnitude of the selection bias depends on the selection process (i.e. assignment mechanism) which consists of the set of rules according to which some members of the population are exposed to the policy while some others are not.

Assuming that the probability to be exposed to the policy depends on a set of characteristics X (of individuals or of the context in which the intervention takes place), if X are unequally distributed among exposed and non-exposed subjects and in turn affect the outcome variable, the selection bias arises. This implies that the outcomes of the two groups would have been different even in the absence of the policy and the observed outcome differences after policy implementation cannot be attributed exclusively to the policy.

The selection bias strictly depends on the rules which determine the exposure to the policy and assign subjects to the treated or control group (selection process). In fact, if subjects were assigned to one of the two groups randomly, both observable and unobservable factors would distribute similarly in the two groups and no bias would arise. Let D, the binary treatment state, be a deterministic function of the triple (X, U, Z), where:

- X is a set of observable characteristics of the units, unaffected by the intervention, possibly correlated to Y^0
- U are unobservable characteristics of the units, unaffected by the intervention, possibly correlated to Y^0
- Z is the observable binary outcome of a random draw (i.e., it is independent of Y^0)

$D(X,U,Z)$ represents the selection process.

In econometric terms the selection bias is caused by the endogeneity of the treatment status. Being the observed individual outcome (Y) a function of the treatment status (D), a set of observable individual characteristics ($\boldsymbol{X}$) and a set of unobservable individual characteristics ($\boldsymbol{U}$):

$$Y = f(D, \boldsymbol{X}, \boldsymbol{U}) \quad (4)$$

if some attributes of *i* (observable or not) simultaneously affect D and Y, the estimate of the average treatment effect obtained by comparing treated and not treated subjects (i.e. the coefficient of D in a multiple regression framework) is affected by selection bias.

Quasi-experimental methods aim at reproducing as closely as possible the key feature of the experimental design: having two groups equivalent in all respects but different with regard to the probability of being exposed to the treatment. Each quasi-experimental method builds on a specific identification strategy which is the way observational data can be used to approximate an experiment.

3. Selection to treatment: a taxonomy

The ability to estimate the impact of an intervention depends on the correct identification of the selection process. In general, three alternative scenarios arise and given the general setup and notation proposed in previous sections, they are summarized as:

- units are randomly assigned to the treatment and control group: random assignment ($D(Z)$ represents the selection process)
- treated and control units differ only with respect to some observable characteristics (X) which in turn affect the outcome: selection on observables ($D(X)$ represents the selection process)
- treated and control units differ with respect to some unobservable characteristics (U) which in turn affect the outcome: selection on observables: selection on unobservables ($D(U)$ represents the selection process)

Even if the selection process strictly depends on the policy structure and its implementation rules, it also depends on data availability. For example, subjects might be exposed to a treatment according to a set of observable characteristics, but no data on those characteristics are available. Thus, the selection is on observables in principles, but the researcher faces a selection on unobservables in practice.

Whether a given selection process can be reasonably assumed for a given treatment requires a case-by-case assessment and ultimately is up to the researcher.

In each of the above scenario, the treatment effect can be identified with specific methods, which might require additional assumptions to return an unbiased estimate of the effect. They will be discussed in what follows.

3.1 Randomized assignment

$\mathrm{D(X, U, Z) = Z}$

When units are assigned randomly to treatment and control group, both observable and unobservable factors are distributed similarly in the two groups, which are equal in expectations. Un unbiased estimate of the treatment effect is obtained by comparing the average outcomes across the treatment and control groups:

$$E(Y^1 \mid D = 1) - E(Y^0 \mid D = 0) = ATT \tag{5}$$

Under this scenario, the ATT equals the Average Treatment Effect (ATE), which is the average effect over the entire sample (treated and untreated units).

3.2 Selection on observables

$\mathrm{D(X, U, Z) = D(X)}$

If treated and control individuals differ only with respect to some observable characteristics (X) which in turn affect the outcome, a proper estimate of ATT can be obtained by simply controlling for those attributes. As a result, the composition of the two groups is made equivalent with respect to the characteristics affecting the outcomes. This can be obtained parametrically (with regression models) or not (with matching techniques), but always requires common support, i.e. for all the possible values of the covariates both treated and no treated units have to be observed.

The condition of selection on observables is also known as *unconfoundedness*, *conditional independence assumption* (CIA), or *ignorable assignment mechanism*. Under the potential outcome notation, given a set of observable covariates X, the potential outcomes are independent of the treatment status:

$$D_i \perp (Y_i^0) | X_i \tag{6}$$

Assumption: selection on observables: all the factors relevant to the outcome and entering the selection process are observable (i.e. the uptake of the program is based entirely on observed characteristics)

Strategy: controlling for the observable attributes affecting the selection process. As a result, the composition of the two groups is made equivalent with respect to the characteristics affecting the outcomes. Once all the observable relevant attributes are controlled for, there are no systematic differences across the treatment and control groups.

Data requirements: post-policy data including measures of the outcome variable for treated and untreated units and a proper set of relevant characteristics (affecting the outcome and the selection process).

Quasi-experimental methods: multiple regression, matching techniques***.***

When applying these methods under a selection of observable scenario, the researcher should pay attention to the risk of using extrapolations to compare incomparable people. To properly identify the causal effect of the treatment, treated and untreated units need to share the same support[3] (*common support* condition) with regard to X (see Figure 1). There is a trade off in the number of observable covariates to use. In fact, the larger the set of reasonable X variables, the more likely to have zero selection bias, however, as the number of X increases, the common support condition may be violated.

Suppose you are interested in assessing the impact of nutrition labels on the quality of diet of consumers. Label readers (the treated group) and label non-readers (non-treated group) systematically differ according to their health orientation, which in turns affect their quality of diet. Suppose you can measure health orientation for the two groups of consumers. Controlling for health-orientation before comparing the diet quality in the two groups is not enough if it is done out of the common support region.

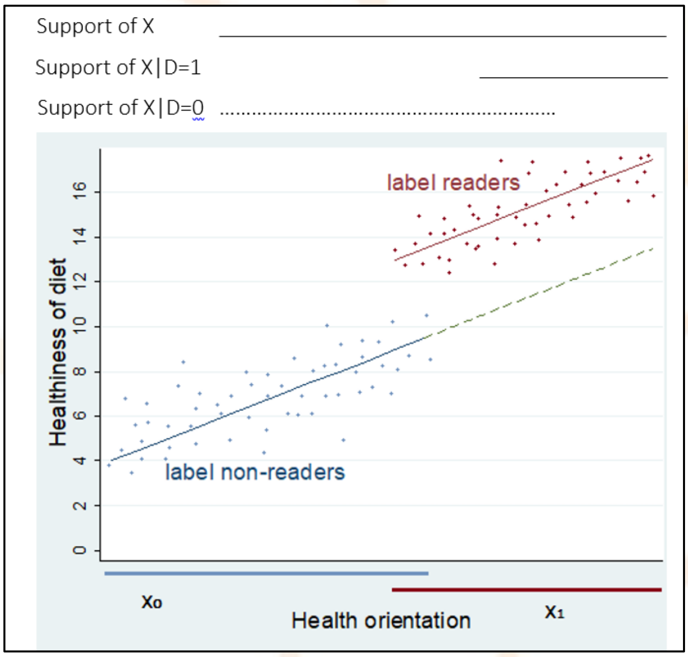


Figure 1 Common support when comparing nutrition labels readers and non-readers

[3] The support of a random variable is the set of realizations that occur with positive probability.

3.2.1 Multiple regression models

Under the hypothesis that all the factors affecting the outcome and entering the selection process are observable (selection on observables), the outcome model for observational data can be written as:

$$y_i = \alpha + \beta D_i + \gamma \boldsymbol{x}_i + \varepsilon_i \quad (7)$$

where y_i is the outcome for the *i-th* unit, D_i is a binary indicator (the *policy dummy*) which is equal to 1 when the *i-th* unit is exposed to the policy and 0 otherwise, $\boldsymbol{x_i}$ is a vector of observed characteristics of the *i-th* unit and β returns a consistent estimate of the ATT. If the assumption of selection on observable holds, the above model can be simply estimated on a cross-section of units observed in a single time period after the policy implementation (no need for pre-policy data).

Application 1. Estimating the effect of the Catalonia soda tax on CPIs using a multiple regression

Dataset: **CPI.dta**, consists in 912 observations: CPIs for non-alcoholic drinks, recorded for 19 autonomous regions on a 48-months-period (from January 2016 to December 2019). Only post-policy observations are used (tot. 380).

Outcome/s: CPIs for non-alcoholic drinks

Treated unit/s: Catalonia

Control unit/s: remaining 18 autonomous regions

Assumption/s: selection on observables. All the factors affecting the outcome and entering the selection process are observable (i.e. treated and control regions differ only according to observable characteristics).

Strategy: estimate the ATT using a multiple linear regression on post-policy data to control for observable characteristics.

$$CPI_i = \alpha + \sum \beta_k x_{ki} + \gamma D_i + u_i \quad (8)$$

Where x_{ki} is a vector of *k* observable characteristics for the *i*-th region, $D_i$is the policy dummy which equals 1 for Catalonia and 0 for the remaining control regions and $u_i$is the idiosyncratic error term.

Under the assumption of selection on observable, γ yields an unbiased estimate of the ATT of the tax.

Table 2 Estimated impact on CPIs. Multiple regression output on post-policy data.

	Monthly CPI (non-alcoholic drinks)
Policy dummy (=1 Catalonya, = rest of the country)	**6.045*****
	(0.454)
Household size	-0.503
	(0.447)
One-person household	8.381***
	(2.560)
One parent with children less than 16 y.o.	9.464***
	(3.321)
One parent with children older than 16	-0.239
	(2.415)
Couple without children	9.210***
	(2.173)
Couple with children less than 16 y.o.	3.612*
	(1.954)
Couple with children older than 16	1.708
	(2.035)
Age of the household reference person	-0.030
	(0.043)
Education level 1	1.818*
	(1.043)
Education level 2	2.839***
	(1.088)
Education level 3	0.535
	(1.427)
Pensioner-only household	-0.760
	(1.307)
Constant	98.585***
	(3.710)
Observations	608
R-squared	0.352

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Application 2. Estimating the effect of the Catalonia soda tax on prices of different drink categories using a multiple regression

Dataset: **PRICE.dta**, consists in 912 observations: 19 autonomous regions on a 48-months-period (from January 2016 to December 2019). Only post-policy observations are used (tot. 380).

Outcome/s: Price of all non-alcoholic drinks, price of non-alcoholic soft-drinks, price of energy drinks, price of sport drinks, price of juices.

Treated unit/s: Catalonia

Control unit/s: remaining 18 autonomous regions

Assumption/s: selection on observables. All the factors affecting the outcome and entering the selection process are observable (i.e. treated and control regions differ only according to observable characteristics).

Strategy: estimate the ATT using a multiple linear regression on post-policy data to control for observable characteristics.

$$Price_{ij} = \alpha + \sum \beta_k x_{ki} + \gamma D_i + u_i \tag{9}$$

Where $Price_{ij}$ is the price of the *k-th* beverage, x_{ki} is a vector of *k* observable characteristics for the *i*-th region, $D_i$is the policy dummy which equals 1 for Catalonia and 0 for the remaining control regions and $u_i$is the idiosyncratic error term.

Under the assumption of selection on observable, γ yields an unbiased estimate of the ATT of the tax on the price of the *k-th* drink.

Table 3 Estimated impact on the price of different drink categories. Multiple regression output on post-policy data.

	PRICE (All non-alcoholic drinks)	PRICE (Soft-drinks)	PRICE (Energy drinks)	PRICE (Sport drinks)	PRICE (Juice)
Policy dummy (=1 Catalunya, = rest of the country)	-0.0474	0.104**	-0.0895	0.0255	0.122**
	(0.0369)	(0.0409)	(0.205)	(0.104)	(0.0560)
Household size	0.00315	-0.0809**	-0.335	0.299	0.103*
	(0.0350)	(0.0392)	(0.345)	(0.188)	(0.0548)
One-person household	0.275	0.0343	-3.520**	1.315	0.764**
	(0.208)	(0.231)	(1.684)	(0.852)	(0.318)
One parent with children less than 16 y.o.	0.208	0.732**	0.0303	2.346**	0.268
	(0.284)	(0.315)	(2.322)	(1.060)	(0.432)
One parent with children older than 16	1.104***	0.637***	-1.733	0.256	0.205
	(0.191)	(0.212)	(1.708)	(0.808)	(0.330)
Couple without children	0.206	-0.0461	-2.812**	0.961	0.743***
	(0.180)	(0.200)	(1.382)	(0.708)	(0.280)
Couple with children less than 16 y.o.	0.272*	0.112	-1.516	0.594	0.422*
	(0.154)	(0.174)	(1.147)	(0.556)	(0.235)
Couple with children older than 16	0.195	0.0741	-2.003	1.381**	0.748***
	(0.166)	(0.184)	(1.344)	(0.673)	(0.253)
Age of the household reference person	-0.000878	0.00116	-0.0207	0.000220	0.00458
	(0.00376)	(0.00417)	(0.0287)	(0.0133)	(0.00582)
Education level 1	-0.354***	-0.338***	-1.649**	-0.121	-0.809***
	(0.0881)	(0.0977)	(0.692)	(0.333)	(0.140)
Education level 2	-0.556***	-0.420***	-0.318	0.447	-0.515***
	(0.0854)	(0.0946)	(0.616)	(0.274)	(0.135)
Education level 3	-0.202*	-0.139	-0.728	0.137	-0.358*
	(0.122)	(0.137)	(0.813)	(0.414)	(0.186)
Pensioner-only household	0.114	0.103	0.721	0.241	0.352**
	(0.102)	(0.114)	(0.789)	(0.389)	(0.163)
Constant	0.564*	1.186***	6.732***	-0.559	0.311
	(0.304)	(0.339)	(2.503)	(1.194)	(0.470)
Observations	380	379	296	339	378
R-squared	0.236	0.219	0.055	0.055	0.194

3.2.2 Matching methods

The class of methods based on matching – propensity score matching (PSM) in particular– has been popular in health sciences, but it is hardly useful without combining it with other quasi-experimental methods. In fact, an effective matching requires full knowledge of the structural model determining outcomes, or full information about the selection process. However, in such situation one could simply obtain the ATT using an OLS regression model as in equation 7**Fehler! Verweisquelle konnte nicht gefunden werden.**. Not only, but authoritative recent studies have emphasized that improper application of PSM could lead to the opposite (and highly undesirable) result of increasing unbalances in unobservables, and lead to larger biases (King & Nielsen, 2019) .

Nevertheless, PSM is widely used, probably because it is an intuitive and relatively easy to teach method, and statistical packages offer fast implementation algorithms. Technical details can be found in Caliendo and Kopeinig 2008. PSM aims at balancing the distribution – or at least the means – of observables between the treated and the control samples. It does so by working on the control sample, by dropping observations, or by applying weights. For example, an observation in the treatment group can be matched with a single observation in the control group, or with a weighted average of observations from the control group. How this matching is accomplished depends on the matching algorithm, and there are many variants: nearest neighbour, radius, kernel and stratification matching being those most commonly implemented. The idea is that rather than matching on the full set of observable variables $\mathbf{x}$ (affecting the outcome and the selection process), a synthetic function of these variables is used, the propensity score. A propensity score is the probability of a unit to end up in the treatment group given its observed characteristics $\mathbf{x}$, and can be easily estimated via a probit or logit model. Matching on the probabilities estimated through these models is easier and more feasible than attempting to match all individual characteristics.

The assumption of unconfoundedness – which requires that no relevant unobservable exists – cannot be tested directly. However, propensity scores are based on the estimation of a binary dependent variable model, and goodness-of-fit measures for that model, e.g. the Pseudo-R^2 or the rate of correct predictions, provide relevant feedback. Even if we find that most of the covariates are relevant (significant) in explaining the assignment-to-treatment process, low goodness-of-fit diagnostics signal that our observables are not enough, and the unconfoundedness assumption is not credible. More sophisticated testing strategies exist, as the Rosenbaum bounds or IV-based tests (see DiPrete & Gangl, 2004), but one should be wary of any PSM studies that does not provide strong evidence that the unconfoundedness assumption is met, as ATT estimates may otherwise be affected by a large bias.

Beyond this, PSM requires overlapping of the propensity scores ranges between the treatment and control group. In a non-random setting we are likely to find higher propensity scores in the target group, and some of them might be too high to find the right match in the control group. In that case, unmatchable observations are dropped from the target group, which means that the estimated ATT does not refer to the original treated sample, but to the reduced one. This might become a major limitation for the ATT estimate. Imagine that in a voluntary food assistance the poorest individuals are very likely to participate, hence have very high propensity scores, but they are not accounted in the ATT estimate because no adequate match is found. Then, the ATT will measure the impact of the policy on a population which excludes those who benefit the most.

Relative to other methods, PSM evaluations are less popular in nutrition policy analysis, but several applications can be found in the literature. Clark & Fox, 2009 apply matching methods to investigate the impact of the US School Breakfast and National School Lunch Programs on vitamin, mineral and sodium intakes. The method seems to be more popular among development economists, for example Abebaw et al., 2010 use PSM to estimate the effects of a food security program in Northwestern Ethiopia.

3.3 Special case of selection on observables: deterministic assignment rule

$\mathrm{D}(\mathrm{X}, \mathrm{U}, \mathrm{Z}) = \mathrm{D}(\mathrm{X})$

In some special cases there is an observable deterministic rule assigning subjects to the treatment. Typically, this is the case when the selection process is driven by administrative rules, e.g.:

$$D_i = I(x_i > x_0) \qquad (10)$$

where individuals whose attribute x_i is higher than a threshold level x_0 participate to the intervention (e.g. people are exposed to the policy if they are older than 14, as for the French vending machine ban, see Capacci, Mazzocchi, and Shankar (2018). Units just on the right of the threshold x_0 and units just on the left of it are equivalent with respect to x (provided x is the only variable driving the selection process, *continuity condition*) and any other characteristic whether observable or not. Thus, comparing the two groups around the threshold gives a good estimate of the average treatment effect for the treated people since around the threshold the assignment to treatment is as good as random (the two groups are approximately equivalent with respect to all the characteristics relevant for the outcome)

Assumptions: selection on observables and *continuity assumption*. When the continuity assumption holds, the potential outcome is a continuous function of the assignment variable x, so that the policy is the only factor affecting the continuity of the outcome function around the threshold.

Strategy: comparing treated and untreated units around the threshold, where thanks to the continuity assumption, the assignment to treatment is as good as random (the two groups are approximately equivalent with respect to all the characteristics relevant for the outcome).

Quasi-experimental methods: Regression Discontinuity Designs (RDD), sharp/fuzzy

Limitations and caveats: this approach to impact estimation suffer from potentially poor external validity. In fact, individuals close to the threshold x_0 might not be representative of the wider population.

3.3.1 Regression Discontinuity Designs

For some specific policies, eligibility depends on the threshold value for a single continuous variable. Typical examples are policies designed around an administrative eligibility criterion based on age or income thresholds to allow access to food assistance programs or other subsidies. When such a sharp classification exists and the variable is known, the division between target and control units is straightforward. As this variable is most likely to be a key determinant for the outcome of interest, this also implies that there is no overlapping and two sub-population are hardly comparable.

In these cases, restricting the analysis to those units that are just below or just above the threshold is a potential solution. With a very large sample, the researcher might have a sufficient number of observations even after restricting the dataset. For example, if a policy is targeting subjects aged below 30, and we have a large data set including individuals within 6 months from their 30th birthday, the resulting sample is relatively homogeneous in terms of age, and splitting the sample in two groups through the date of birth is similar to randomized assignment, and one should not expect major selection biases. A mere mean comparison test between the average outcomes could be a quite good estimate of the ATT.

However, one major caveat accompanies this estimate of the treatment effect, which is certainly valid in the selected neighbourhood of the cut-off point, but not necessarily for data points further away. In our example, we may get good and reliable estimates of the ATT for those aged 30, but we can say little about the policy effects on those that are aged 20 or 25 relative to those aged 35 or 40. Thus, ATTs estimated through RDD are characterized by limited external validity.

Furthermore, this threshold analysis commonly runs into two major issues: (1) the number of available observations around the cut-off value is not large; (2) the cut-off point may be associated with a number of confounding events creating discontinuities. For example, if the age cut-off is also the retirement age (e.g. 65), one may think that such an event creates relevant disparities between the target and control groups in variables that may in turn affect the outcome variable.

The first problem is addressed by relying on the functional relationship between the outcome and the assignment (running) variable. When such function is identifiable, it can be exploited to expand the sample of interest. To do that, we need to assume continuity, which means that without the policy the outcome would just follow the identified functional relationship with the running variable. The most basic functional form is a simple bivariate linear regression, and the policy impact would be captured by a sharp shift in the intercept as the running variable reaches the cut-off point. By exploiting this linear relationship, one is able to expand the sample and consider units that are further away from the threshold. This brings in a second assumption, linearity, which requires that the linear relationship is valid within the expanded neighbourhood of the cut-off point. Although few relationships between the outcome and the running variable are indeed linear, when the neighbourhood under consideration is still relatively small, then the linear approximation performs well and the ATT estimate becomes more credible (and efficient) for the sample of interest. In other words, its internal validity is higher. Clearly this introduces a trade-off between internal validity and efficiency. If we use a large neighbourhood, we have more observations and a more efficient estimate of the ATT. However, observations are more heterogeneous, the linearity assumption becomes more influential, and there is less internal validity.

RDD deals well with unobservables when these are unlikely to differ substantially between the two groups within a small neighbourhood of the cut-off point. However, the crucial continuity assumption implies that there are no other major “jumps” in relevant outcome determinants at the same cut-off point. There are cases when this assumption is clearly challenged, for example when the cut-off value is one with administrative and legal relevance. For example, age cut-offs at 18 and 65 are common to several economic and health policy measures, or some income eligibility threshold levels can be similar across different policies in the same country, which complicates the attribution of the causal effect to a specific policy. In such cases, the only viable solution seems to be the inclusion in the model of covariates which help to control the confounding effects (Frölich & Huber, 2019). More generally, one should test whether the continuity assumption holds simply by applying the same RDD model on relevant confounding factors, and expecting not to find significant discontinuities. One key reason why the continuity assumption fails to hold is when subjects have some control on the assignment variable. For example, one might delay some revenue (job offer) to maintain eligibility for a program based on income thresholds. If these behaviours (“bunching”) are possible, then the continuity assumption is challenged and RDD becomes less credible[4].

Since the estimation of causal effects through RDD depends on assumption on the neighbourhood size and the shape of the relationship between the outcome and the running variable, a number of extensions and variants in the estimation procedures exist. First, the optimal size of the window around the cut-off point (the bandwidth) may be also an output of the estimation algorithm. Second, non-parametric regressions allow to relax the assumption of a linear relationship, and place different weights on observations depending on how far they are from the cut-off point. Third, when the running variable does not determine a sharp cut-off (i.e. all individuals meeting the rule are treated), but only creates a shift in the probability to be treated, then fuzzy RDD better serves for the purpose. This is the case of voluntary policies, where not all eligible individuals are exposed, and/or when there are exception allowing participation of subjects that do not meet the cut-off eligibility requirement.

Including covariates, changing the bandwidth, allowing for non-linear relationships, or opting for a fuzzy design are all choices that may potentially lead to different result, which is why convincing robustness checks are not an optional feature for RDD studies. On the one hand, one may want to show that the estimate of the causal effect is relatively consistent across different choices. On the other hand, falsification tests add credibility to the identification strategy. For example, one may want to show that different cut-off points other than the one relevant to the analysis are not associated with discontinuities.

Although the range of policies that are suitable to this method is limited, and the aforementioned external validity caveat applies, RDD is considered a relatively powerful causal identification method. Sometime researchers have expanded the scope of RDD by considering time as the assignment variable with panel or time series data (see e.g. Aguilar et al. 2021) . In these exercises, the idea is that comparing outcome just before and after the time of the policy implementation, while exploiting some outcome-time relationship, may lead to the identification of the policy causal effect. However, this also

[4] Interestingly, this opens the way to relevant behavioral evaluations and estimation which exploit the possibility to identify manipulation (see Kleven 2016).

leads to major differences in the requirements for successful identification relative to the standard RDD method, an issue which deserves careful consideration before one chooses “time” RDD over simpler event study models (Hausman & Rapson, 2018).

Examples of RDD application to nutrition policies include the income-eligibility rule for the US School Lunch Program (Schanzenbach, 2009), the removal of vending machines from secondary schools in France (Capacci et al., 2018), the impact on nutrition and wellbeing of a new refugee assistance program in Kenya (MacPherson & Sterck, 2021) , and the effects of microcredit on children nutrition in China (You, 2013).

3.4 Selection on unobservables and time invariant selection bias

$\mathrm{D}(\mathrm{X}, \mathrm{U}, \mathrm{Z}) \ = \ \mathrm{D}(\mathrm{U})$, with U dependent on Y^0, but U independent of the variation of Y^0 over time.

Very often the variables responsible for the selection bias are not observable. The treated and not treated subjects might differ for some latent (or simply not-observed) attributes which might affect their responsiveness to the treatment. For example, this might occur when units self-select to the intervention (See nutrition labels use). In this case those latent variables (as motivation) affecting participation might also be responsible for specific levels of the potential outcome of treated subjects and emphasize the effect of the treatment. Yet, if the outcome is a repeatable event, observed before and after the intervention, and it can be reasonably assumed that the difference between the treated and control group is stable in time (*common* or *parallel trend assumption*), the Difference in Differences (DID) strategy can neutralize the effect of the bias caused by unobservable heterogeneity. This is one of the most exploited quasi-experimental strategies for the evaluation of public interventions in non-experimental setting and consists in comparing the outcome of treated and untreated subjects before and after the treatment (difference in differences).

Assumption: those factors affecting the selectin process are not observable (or observed) but the outcome difference between treated and control subjects in absence of the treatment (i.e. the selection bias, $(\boldsymbol{Y^0} \big| \boldsymbol{D} = \boldsymbol{0}) - \boldsymbol{E}(\boldsymbol{Y^0} \,|\boldsymbol{D} = \boldsymbol{1}))$ is time invariant. This is known as *parallel (or common) trend* condition (see *Figure 2*).

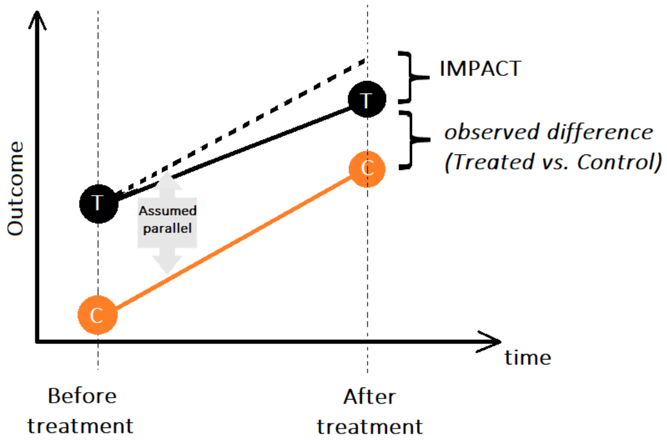


Figure 2 Parallel trend assumption in the DID framework: the selection bias is stable over time

Strategy*:* the difference between the change over time (post policy vs pre policy) for the treated subjects and the change over time (post policy vs pre policy) for the untreated subjects retrieves an unbiased estimate of the ATT.

Data requirements: the outcome $\boldsymbol{Y}$ is a repeatable event and it is observed both before and after the intervention for participants and non-participants. Both multiple cross-sections (the composition of the two group changes in different time periods) and panel data can be used.

Quasi-experimental method: DID multiple regression for multiple-cross sections, DID panel data models

Application 3. Estimating the effect of the Catalonia soda tax on CPIs, using a DID approach and assuming common trend

Dataset: CPI.dta, consists in 912 observations: CPIs for non-alcoholic drinks, recorded for 19 autonomous regions on a 48-months-period (from January 2016 to December 2019).

Outcome*:* CPIs for non-alcoholic drinks

Treated unit/s: Catalonia

Control unit/s: remaining 18 autonomous regions

Strategy: the difference between the average observed change over time for the treated and the average observed change over time for the controls retrieves the ATT. This can be simply obtained by calculating four averages, as in Table 4, where the difference in differences is reported in bold (6.13 represents the estimated increase in CPIs, due to the tax, under the assumption of common trend).

Table 4 Observed average CPIs for non-alcoholic drinks in Catalonia and control regions, calculated before and after the tax.

	Treated region (Catalonia)	**Control region (All other regions)**	**Treated-control**
Pre tax	100.00	99.89	0.10
Post tax	108.17	101.94	6.23
Post-pre	8.17	2.04	**6.13**

The same point estimate of the ATT under the DID framework is easily obtained by estimating the following regression model:

$$y_i = \beta_0 + \beta_1 post_i + \beta_2 treated_i + \gamma\, post_i * treated_i + u \qquad 11$$

where y_i is the outcome for the *i*-th observation, $post_i$ is a binary variable=1 if *i* is observed after the policy, $treated_i$ is a binary variable=1 if *i* is exposed to the policy (the policy dummy) and γ – the coefficient of the interaction term ($post * treated$) – retrieves an unbiased estimate of ATT. The regression outputs quickly return the standard error of the estimated ATT.

The minimum data requirement for a DID estimation is two cross-sections (observed before and after the policy introduction) of treated and control units. However, potential differences in the composition of the treated and control groups might work against the reliability of the common trend assumption. If a panel dataset is available, the DID model can be rearranged as follows:

$$y_{it} = \alpha_i + \lambda_t + \boldsymbol{\delta} T_{it} + \eta_{it} \qquad 12$$

where y_{it} is the outcome observed for unit i at time t,α_i and λ_t are individual and time fixed effects, T_{it} is a binary variable=1 for treated units observed after the introduction of the treatment and $\boldsymbol{\delta}$ returns the estimated ATT. Units and time fixed-effects capture constant outcome differences across units and temporal changes in the outcome that are common to all units. For this reason, they replace the single $treated$ and $post$ dummies respectively. The model can be estimated introducing a set of dummies indicating each time period in the sample and a set of dummies for every unit in the sample.

We apply the standard DID approach and exploit the longitudinal structure of the data to obtain an estimate of the causal effect of the tax on CPIs, assuming common trend and controlling for potential seasonal effects:

$$CPI_{it} = \alpha_0 + \sum_i \gamma_i \, d.region_i + \sum_t \beta_t d.month_t + \delta T_{it} + \boldsymbol{\theta}_{it} + \eta_{it} \tag{13}$$

where CPI_{it} is the CPI for the i-th region in t-th month, $\sum_i d.region_i$ is a set of dummy for every region in the sample (except one, to escape the dummy trap), $\sum_t d.month_t$ is a set of dummy for every month in the sample (except one), T_{it} is a binary variable=1 for Catalonia CPIs registered after May 2017 and $\boldsymbol{\theta}_{it}$ are quarterly dummies to capture seasonal effects. The two sets of regional and monthly indicators control for regional and monthly fixed effects, respectively. δ returns the estimated ATT. In order to estimate δ, Equation 13 is equivalent to the following model:

$$CPI_{it} = \alpha_0 + \alpha_1 treated_{\mathrm{i}} + \alpha_2 post_t + \delta T_{it} + \boldsymbol{\theta}_{it} + \eta_{it} \tag{14}$$

where regional and monthly fixed effects are replaced by the policy dummy $treated_i$ (which equals 1 for Catalonia and 0 for the remaining regions) and $post_t$(which equals 1 when the CPI is observed after May 2017 and 0 otherwise), and T_{it} is in fact equivalent to the interaction term $treated_i * post_t$. Table 5 shows the regression output, with the estimated impact on CPIs in bold.

Table 5 DID estimate of the effect of the Catalonia tax on CPIs using a regression model with fixed effects and quarterly dummies to control for seasonality, common trend assumed.

	Monthly CPI
	(non-alcoholic drinks)
County fixed effects:	
Aragona	-0.738***
	(0.266)
Asturias	-1.145***
	(0.266)
Balearic	2.902***
	(0.266)
Canary	0.0741
	(0.266)
…	
Dummy for quarter=2	3.359***
	(0.423)
Dummy for quarter=3	3.392***
	(0.423)
Dummy for quarter=4	3.329***
	(0.423)
Month fixed effects:	
Month_id=2	0.265
	(0.422)
Month_id=3	0.536
	(0.422)
Month_id=4	-2.900***
	(0.423)
Month_id= 5	-2.878***
	(0.423)
…	
Interaction term: treated*post	**6.130*** **
	(0.409)
Constant	99.83***
	(0.350)
Observations	912
R-squared	0.794

3.4.1 Is the selection bias time invariant? Testing the common trend assumption

The common trend assumption is crucial under the DID approach and requires that the difference in the outcome between treated and control units with no treatment (the selection bias) is time invariant. When the available sample consists in multiple units observed over multiple time periods. It is possible to investigate outcomes' time trends among the treated and control groups and provide evidence on the appropriateness (or not) of the common trend assumption. In the present section, we discuss three possible strategies to do so.

A fist strategy to test whether the common trend assumption holds is to estimate a pre-policy model including a trend variable interacted with the policy dummy. This allows to check whether before the policy introduction a systematic difference in trends exists between treated and control units. The model can be specified as follows:

$$y_{it} = \alpha_0 + \alpha_1 trend_t + \alpha_2 treated_{\mathrm{i}} + \mathrm{\alpha}_3 trend_{\mathrm{t}} * treated_{\mathrm{i}} + \mathrm{u}_{\mathrm{it}} \qquad 15$$

where y_{it} is the outcome observed for the i-th unit at time t before the policy introduction, $trend$ represents a linear time trend, $treated$ is the policy dummy (=1 if the unit belongs to the treatment group and =0 otherwise) and α_3 returns an estimate of the expected difference in the linear trend between the treated and the control group before the policy was implemented.

A second strategy consists in the visual inspection of the data. By plotting average outcomes for treated and control units against time it is possible to get a sense of whether the common trend assumption is plausible or not.

A third strategy consists in estimating a model for the outcome variable over the entire sample of observations (pre and post) against a full set of time dummies interacted with the policy dummy:

$$y_{it} = \alpha_0 + \sum \gamma_k d.time_k + \sum \beta_k d.time_k treated_i + u_{it} \qquad 16$$

where $d.time_k$ is a dummy for the k-th time period and $treated_i$ is the usual policy indicator for the i-*th* unit. The coefficients of the k interaction terms will tell if systematic differences in time fixed effects exist between treated and control units.

All the above strategies are applied to the Catalonia case study in the following section.

Application 4. Testing the common trend assumption for CPIs in the treated and untreated regions

Dataset: CPI.dta

Strategies:

1. testing the hypothesis of common linear time trend between treated and untreated regions using a regression model on pre-policy data. A linear time trend term is interacted with a policy dummy to check for potential differences in the (linear) trend. A significant interaction term will reveal significant differences in linear trends before the tax and should prevent from assuming common trend when estimating the tax effect.

2. Plotting average outcomes for the treated and control groups against time give a sense of potential different time trends.

3. estimating a model for the outcome variable over the entire sample of observations (both pre and post) against a full set of time dummies interacted with the policy dummy to detect systematic differences in time fixed effects among the two groups.

Strategy 1: Equation 15 is estimated on CPIs over the pre-tax sample and regression outputs are shown in

Table 6. The coefficient of the interaction term (α_3 in equation 15) is statistically significant and positive (in bold below), showing significantly different linear trends in Catalonia and in the rest of the country before the tax was implemented.

Table 6. CPIs regression model estimated on pre-tax observations. Interaction term included to test differential linear time trend between Catalonia and control regions.

	Monthly CPI (non-alcoholic drinks)
Linear trend	-0.0751***
	(0.0118)
Policy dummy (=1 Catalunya, = 0 rest of the country)	-1.025**
	(0.497)
Linear trend*Policy dummy	**0.133****
	(0.0514)
Constant	100.5***
	(0.114)
Observations	304
R-squared	0.123

Strategy 2: Visual inspection is helpful in detecting potential differential time trends. In Figure 3 the Catalonian and the National CPIs are plotted against. The graph shows a steeper linear time trend for CPIs measured in Catalonia with respect to the National CPIs before May 2017, when the tax was implemented.

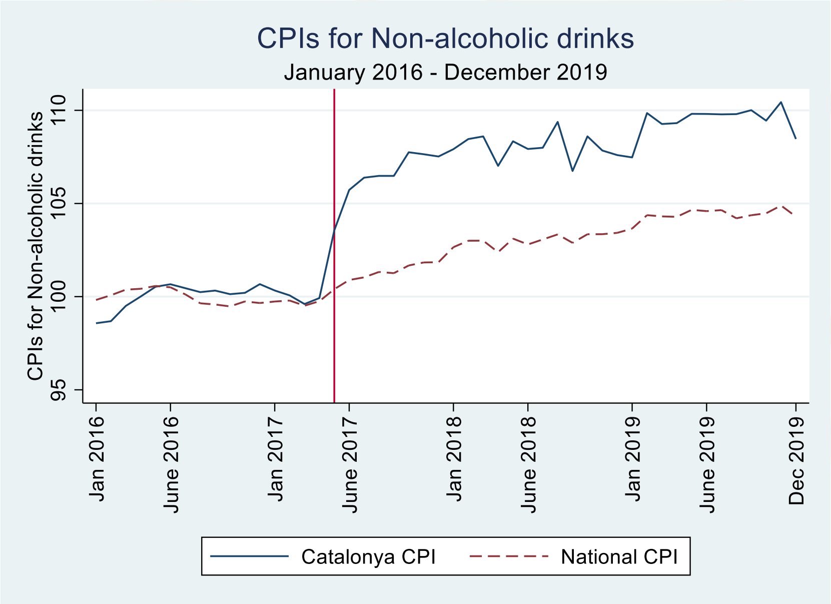


Figure 3: National and Catalonian monthly CPIs for non-alcoholic drinks, January 2016 – December 2019.

The above results sound as a red flag when assuming common trends for DID methods.

Strategy 3: Equation 16 is estimated for CPIs over the fulltime span. Results reported in Table 7 show non-significant interaction terms, nor before neither after the tax. However, when estimated coefficients are plotted against time a systematic pattern emerges (Figure 4). The estimated coefficients are systematically below zero before the tax and above zero after the tax. This pattern suggests some effect of the tax.

Table 7 CPIs regression model with full set of time dummies interacted with the policy dummy, entire sample.

			CPI		
Monthly fixed-effects:					
1.month_id	0.513	1.treated#18.month_id	1.979	1.treated#39.month_id	1.857
	(0.658)		(2.943)		(2.943)
2.month_id	0.786	1.treated#19.month_id	2.555	1.treated#40.month_id	1.963
	(0.658)		(2.943)		(2.943)
3.month_id	1.028	1.treated#20.month_id	2.363	1.treated#41.month_id	2.290
	(0.658)		(2.943)		(2.943)
…		1.treated#21.month_id	2.386	1.treated#42.month_id	2.393
Treated	3.965*		(2.943)		(2.943)
	(2.081)	1.treated#22.month_id	3.363	1.treated#43.month_id	2.210
1.treated#1.month_id	-5.499*		(2.943)		(2.943)
	(2.943)	1.treated#23.month_id	3.078	1.treated#44.month_id	2.879
1.treated#2.month_id	-5.661*		(2.943)		(2.943)
	(2.943)	1.treated#24.month_id	3.148	1.treated#45.month_id	2.585
1.treated#3.month_id	-5.084*		(2.943)		(2.943)
	(2.943)	1.treated#25.month_id	2.718	1.treated#46.month_id	2.067
1.treated#4.month_id	-4.479		(2.943)		(2.943)
	(2.943)	1.treated#26.month_id	2.832	1.treated#47.month_id	2.672
1.treated#5.month_id	-3.958		(2.943)		(2.943)
	(2.943)	1.treated#27.month_id	2.918	1.treated#48.month_id	1.024
1.treated#6.month_id	-3.544		(2.943)		(2.943)
	(2.943)	1.treated#28.month_id	1.811		
1.treated#7.month_id	-3.360		(2.943)	Constant	99.59***
	(2.943)	1.treated#29.month_id	2.533		(0.465)
1.treated#8.month_id	-3.219		(2.943)		
	(2.943)	1.treated#30.month_id	2.366	Observations	960
1.treated#9.month_id	-3.178		(2.943)	R-squared	0.501
	(2.943)	1.treated#31.month_id	2.022		
1.treated#10.month_id	-3.219		(2.943)		
	(2.943)	1.treated#32.month_id	3.362		
1.treated#11.month_id	-3.453		(2.943)		
	(2.943)	1.treated#33.month_id	0.771		
1.treated#12.month_id	-2.928		(2.943)		
	(2.943)	1.treated#34.month_id	2.341		
1.treated#13.month_id	-3.448		(2.943)		
	(2.943)	1.treated#35.month_id	1.504		
1.treated#14.month_id	-3.548		(2.943)		
	(2.943)	1.treated#36.month_id	1.203		
1.treated#15.month_id	-3.784		(2.943)		
	(2.943)	1.treated#37.month_id	0.852		
1.treated#16.month_id	-3.459		(2.943)		
	(2.943)	1.treated#38.month_id	2.559		
1o.treated#17b.month_id	0		(2.943)		

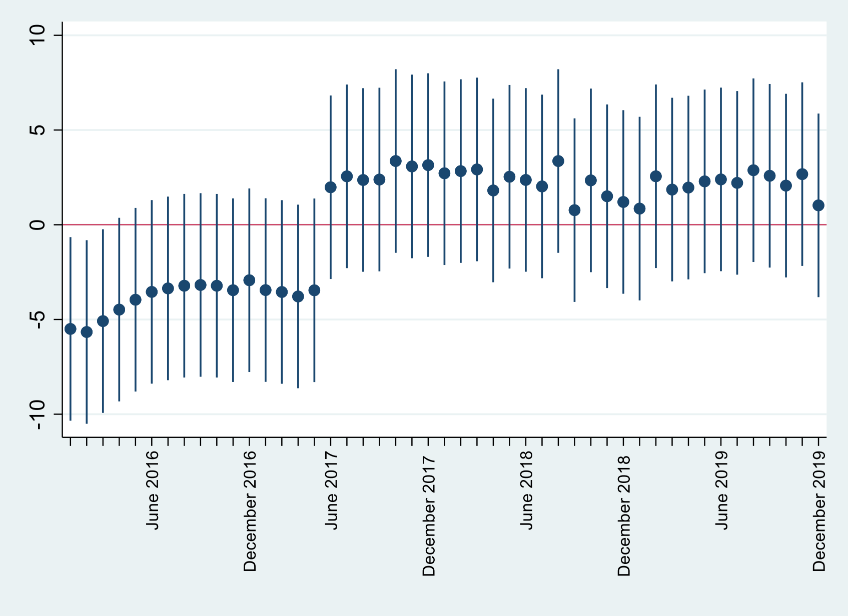


Figure 4 Estimated coefficients of full-set of interaction terms plotted against time.

3.4.2 Accounting for differential (linear) time trends under a DID framework.

When the available sample includes multiple units and multiple time periods, it is possible to slightly relax the common trend assumption and introduce a "degree of nonparallel evolution in outcomes"(Angrist & Pischke, 2014) between units in the absence of the treatment effect. The following model controls for group-specific linear trends:

$$y_{it} = \alpha_0 + \alpha_1 treated_i + \alpha_2 post_t + \beta_1 \mathrm{trend_t} + \beta_2 trend_{\mathrm{t}} * treated_{\mathrm{i}} + \delta T_{it} + \boldsymbol{\theta}_{it} + \eta_{it} \tag{17}$$

Where the standard DID specification (equation 12 and 13) is augmented with a linear trend ($trend_t$) interacted with the policy dummy ($trend_{\mathrm{t}} * treated_{\mathrm{i}}$) to account for differential linear trends in treated and untreated outcomes. This model assumes that in the absence of the policy effect, the outcome in the treated group deviate from the common trend and follow a different (still linear) time trend captured by the β_2 coefficient. δ in equation 17 returns the estimated policy impact (ATT), which consists in "deviations from otherwise smooth trends, even if trends are not common" (Angrist & Pischke, 2014).

Application 5. Estimating the effect of the Catalonia soda tax on CPIs, using a DID approach and allowing for differential linear time trend

Dataset: CPI.dta

Strategy: following equation 17, we estimate the following model:

$$CPI_{it} = \alpha_0 + \alpha_1 treated_{\mathrm{i}} + \alpha_2 post_t + \beta_1 \mathrm{trend_t} + \beta_2 trend_{\mathrm{t}} * treated_{\mathrm{i}} + \delta T_{it} + \boldsymbol{\theta}_{it} + \eta_{it}$$

Where the standard DID model is augmented with a linear trend ($trend_t$) interacted with the policy dummy ($trend_{\mathrm{t}} * treated_{\mathrm{i}}$) to account for differential linear trends in CPIs in Catalonia and the rest of the country. Results are reported in Table 8 and the estimated ATT is in bold. Allowing for potential differential linear time trend slightly reduces the size of the tax effect.

Table 8 DID estimate of the effect of the Catalonia tax on CPIs using a regression model accounting for seasonality and differential linear time trends.

	Montlhy CPI (non-alcoholic drinks)
Linear trend	0.118***
	(0.00891)
Linear trend*policy dummy	0.00649
	(0.0382)
Policy dummy (=1 Catalunya, = 0 rest of the country)	0.0480
	(0.620)
Post policy dummy (=1 if observed after May 2017)	-0.692***
	(0.260)
Interaction term: treated*post	**5.974*** **
	(1.122)
Dummy for quarter=1	0.555***
	(0.198)
Dummy for quarter=2	0.401**
	(0.196)
Dummy for quarter=3	0.121
	(0.195)
Constant	98.56***
	(0.204)
Observations	912
R-squared	0.449

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Application 6. The DID model applied to Case study 2: the Cycling May Campaign

Within Sub task 3.1.3 the DID approach has been deeply investigated also with reference to Case Study 2 which focuses on the Cycling May campaign. In the present section the case study is shown briefly. Thanks to the fruitful collaboration between WP3 (Sub task 3.1.3) and WP6 on this case study, a paper with title "Impact evaluation of a cycling promotion campaign using daily bicycle counters data: the case of Cycling May in Poland" was produced and submitted by Beatrice Biondi (University of Bologna), Aleksandra Romanowska (Gdansk University of Technology) and Krystian Birr (Gdansk University of Technology) on January 2022.

Startegy: to estimate the effect of Cycling May campaign on bicycle traffic in Gdansk we use a difference-in-differences approach by comparing bicycle traffic in Gdansk – where the intervention was implemented – with bicycle traffic in Lodz where Cycling May was not implemented. We estimate panel regression models that account for fixed cross-sectional (counter-level) and time effects. The estimated model has the following equation:

$$y_{it} \; = \; \gamma \, X_{ct} + \delta \, p_{it} + \beta \, Z_t + \; \lambda_i + \varepsilon_{it} \qquad (18)$$

The dependent variable is the number of bicycles registered by each counter i on day t. X_{ct} are city and time specific variables, Z_t are time specific variables, λ_i are counter-specific fixed effects, ε_{it} is the residual error component. The binary variable p_{it} refers to the policy and is 1 for May observations from Gdansk counters, and 0 otherwise, so that the coefficient δ is the DiD estimator of the average treatment effect on treated units (ATT). This coefficient returns the additional bicycle traffic generated in Gdansk by the Cycling May intervention, after controlling for covariates.

X_{ct} in equation 18 are city and time specific control variables that account for specific weather conditions and touristic flows in each city:

$$X_{ct} = \gamma_1 Tourists_{cm} + \gamma_2 WindSpeed_{ct} + \gamma_3 Cloudiness_{ct} + \gamma_4 Temperature_{ct} + \gamma_5 Precipitations_{ct} \qquad (19)$$

The $Tourists_{cm}$ variable refers to the monthly number of tourists in Gdansk and in Lodz, $\forall \, day \; t \in \; month = m$, $with \;\; m = 1, \dots, 36$. Daily average wind speed, cloudiness, temperature and precipitation are included for each city.

The vector Z_t in equation 18 contains time specific variables that are common to the treated and control group, including seasonal factors relevant to commuting, i.e. differences in bicycle traffic by quarter m, day of the week and holidays in Poland. We specify the Z_t component of the model as follows:

$$Z_t = \sum_{m=1}^{3} \beta_{1m} Quarter_m + \sum_{d=1}^{6} \beta_{2t} DayofWeek_t + \beta_3 PublicHolidays_t + \beta_4 SummerHolidays_t \tag{20}$$

The counter-specific fixed effects included in the model (λ_i in equation 18) allow to control for all time-invariant differences between bike lanes, e.g. distance from schools, average traffic, etc. By including these fixed effects, we control for any omitted time-invariant factor. ε_{it} is an error term that captures any unobserved factor that may affect bicycle traffic and is assumed to have zero mean, conditional on the counter and day.

By including exogenous time-varying controls into the model, the common trend assumption between the two groups is met conditionally on these covariates. In the current application, the common trend assumption can be tested by comparing the outcomes over months in which the intervention is not implemented (non-treatment periods), without and with control variables.

The identifying assumption of our model is that there are no omitted factors that have a differential impact on bicycle traffic in Gdanks and in Lodz. In other words, we assume that, after controlling for systematic differences through fixed effects and for city and time-varying differences due to observed factors (atmospheric conditions and touristic flows), in absence of the policy the average outcome in the two cities would be the same. We provide evidence that this assumption is reasonable by showing trends and seasonality in bicycle traffic in the two cities, and the distribution of touristic flow over time.

In order to test for the robustness of our findings, we estimate several models that differ in term of choice of the dependent variable, specification and estimation method, i.e. (1) fixed effects panel regression with the natural logarithm of daily bicycle counts as dependent variable (therefore coefficients estimates are to be interpreted as percentage changes in the number of bycicles); (2) panel Poisson regression using the level of the count, here estimates refer to the difference between the log of expected counts and their exponential function represents the change in the expected number of bicycles (3) Cycling May intervention included as number of yearly participants (instead of as binary variable).

A similar DID model applied to daily bicycle counter data has been used before in the study by (Kraus & Koch, 2021). While we partially follow their approach, the evaluation of Cycling May campaign has one peculiarity related to the fact that the intervention is repeated over time. This means that there are no unique pre- and post- implementation periods, but treatment periods (the months of May) and non-treatment periods are cyclic. The effect of the policy is observable during the month of May, but we expect that there might be a persistence effect in the subsequent period. To explore such persistence, we check for differential outcomes for the months of June, July and August.

Data: the data cover a three years period, from the beginning of September 2016 to the end of August 2019. Cycling May campaign was implemented in Gdansk throughout the period of our analysis (in the month of May for all years from 2017 to 2019). The raw data include daily bicycle counts from 26 counters in Gdansk and 9 counters in Lodz. Observations from three counters in Gdansk were excluded because these counters were installed after the beginning of our period of analysis; observations from two counters in Lodz were also excluded from the analysis because of missing data (likely due to counters technical problems). Therefore, our dataset includes bicycle counts from 30 counters (23 in Gdansk, 7 in Lodz) over a period of 1.095 days, resulting in 32.850 observations.

Outcome: The daily bicycle traffic at the counter level is our outcome of interest. Figure 5 shows the average (across counters) number of daily bicycles counts in the two cities, by month and year. The trend is similar, except for the months from May to August, when the bicycle traffic in Gdansk is relatively more intense. The introduction of the bike sharing system in Gdansk in April 2019 produces a clear increase in bicycle traffic.

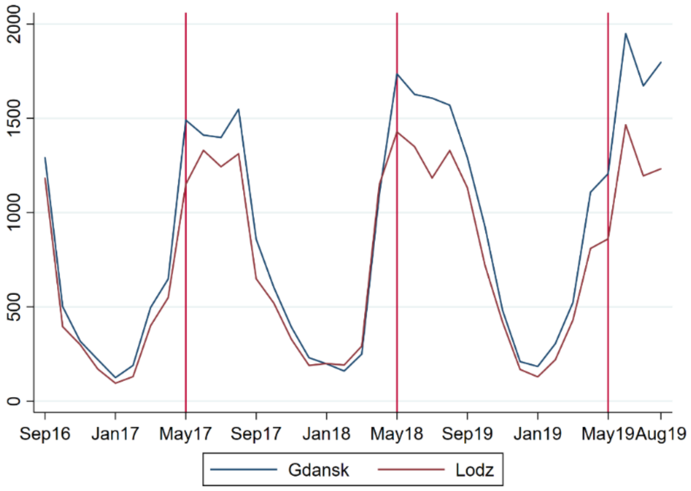


Figure 5 Average number of bicycle counts by city and month

A possible explanation for the gap in bicycle traffic between the two cities during the summer is related to the tourist flows; since Gdansk is situated by the sea, tourist numbers are likely to escalate during the summer.

Figure 6 displays the monthly number of tourists in the two cities, and clearly shows that Gdansk is characterized by highly seasonal tourism, while tourism flows in Lodz are more stable across different seasons and during the period considered. Accordingly, the number of tourists could become a good control in the model, accounting for the different bicycle traffic during the summer.

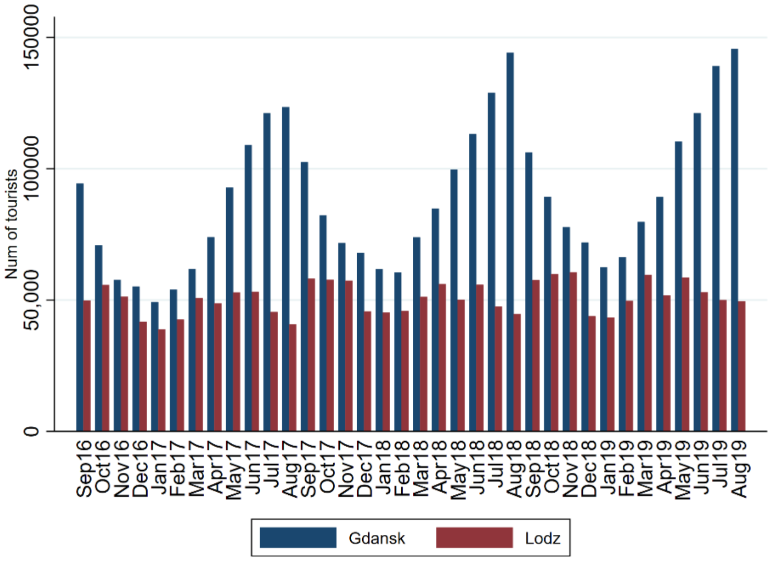


Figure 6 Average number of daily tourists, trend by month and city

Other included covariates (i.e. controlled factor) in the DID model are daily weather data obtained from meteorological stations located in Gdansk and Lodz, close to the central areas, regarding average temperature, average wind speed, average cloudiness and sum of precipitation. In general, Gdansk and Lodz have similar temperature, cloudiness and precipitations, but Gdansk seems windier than Lodz, probably because of proximity to the sea. Focusing on the month of May – the treatment period – Figure 7 shows that, differently from what observed in previous years, precipitations in

May 2019 are substantially higher in Gdansk than in Lodz. Lastly, in the model we control for the period in which bike sharing was active in Gdansk, that is after April 2019.

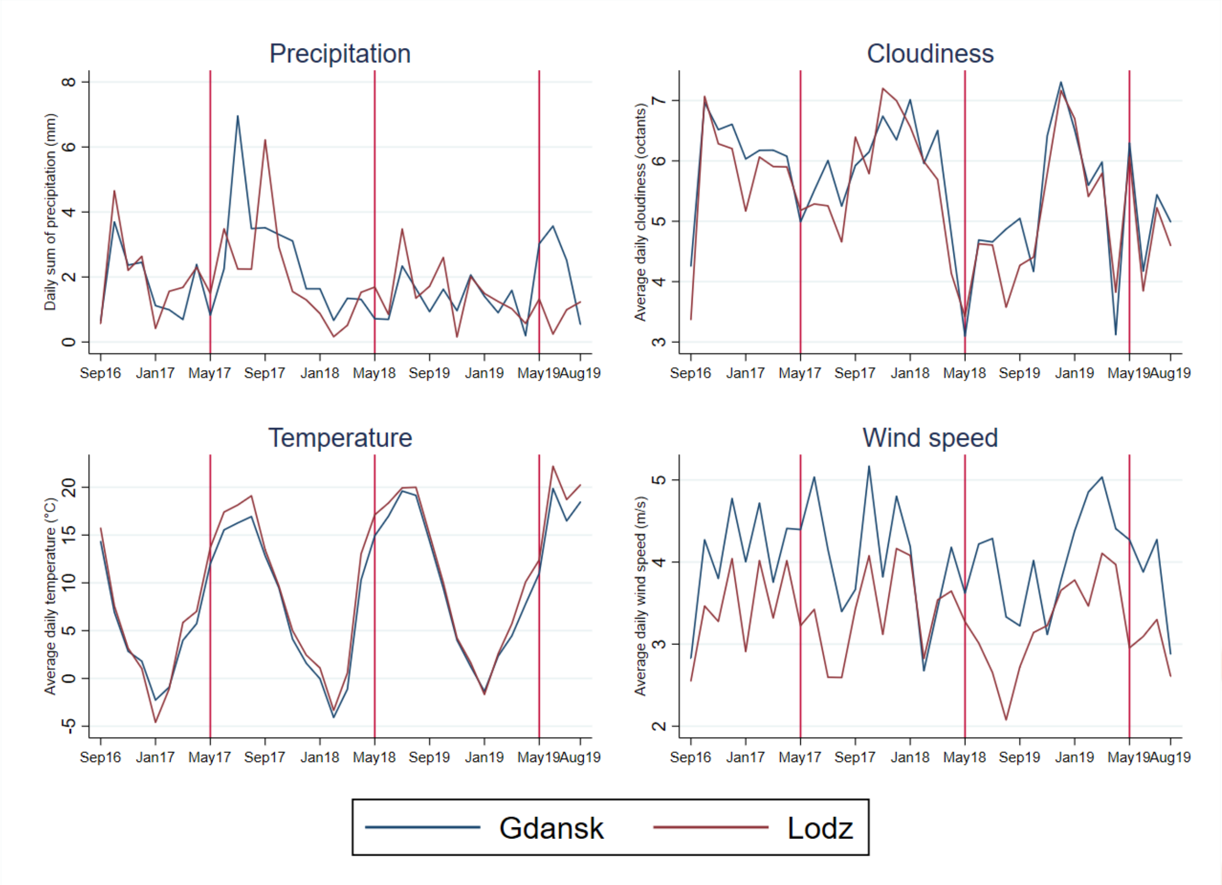


Figure 7 Wheather conditions in Gdanks and Lodz over the estimation sample

We test the common trend assumption outside the intervention periods, i.e. excluding May. More specifically, we allow for differential linear trends between the two cities. Table 9 shows results of panel regression models estimated on daily bicycle counts as dependent variable: model (1) checks for the common trend assumption over time without controlling for any other exogenous factor and model (2) includes potentially relevant covariates. Model 1 results in a significant differential trend, that becomes insignificant when controlling for exogenous factors; also, the common trend becomes smaller and negative, meaning that the positive trend captured by model (1) can be explained but other exogenous factors. Therefore, common trend assumption holds after controlling for seasonality, number of tourists, day of the week, holidays, atmospheric conditions, and activation of bike sharing in Gdansk, and we control for these factors in the DiD model.

Table 9 Check for common trend in non-treatment periods, without controls (1) and including controls (2)

	(1)		(2)	
Trend	11.92***	(0.97)	-2.52***	(0.69)
Trend # Gdansk	8.20***	(1.11)	0.44	(0.85)
Wind speed			-32.59***	(1.75)
Cloudiness			-53.77***	(1.61)
Temperature			42.28***	(0.73)
Precipitation			-13.97***	(0.62)
Number of tourists (thous.)			5.53***	(0.25)
Monday			132.10***	(11.52)
Tuesday			172.89***	(11.56)
Wednesday			167.39***	(11.61)
Thursday			154.50***	(11.53)
Friday			94.93***	(11.53)
Saturday			-22.36*	(11.51)
Public holiday			-150.83***	(18.03)
Summer holiday			111.08***	(13.72)
Quarter=2			207.26***	(13.92)
Quarter=3			6.37	(16.54)
Quarter=4			-51.04***	(9.15)
Bike sharing Gdansk			147.19***	(14.29)
Constant	483.29***	(9.18)	338.38***	(20.80)
N		30,060		30,060
Log likelihood		-242727		-231430

Standard errors in parentheses. * p<0.1; ** p<0.05; *** p<0.01

Results of the DID model are displayed in *Table 10* (model a): according to the estimated model, Cycling May campaign in Gdansk increases bicycle traffic by nearly 159 bicycles per day per counter, ceteris paribus.

Table 10 Estimated effect of the Cycling May campaign using a DID approach

Model (a)	Coefficient	Standard error
Wind speed	-33.67***	(1.74)
Cloudiness	-59.90***	(1.54)
Temperature	43.72***	(0.71)
Precipitation	-14.49***	(0.63)
Number of tourists (th.)	5.13***	(0.24)
Monday	133.72***	(11.32)
Tuesday	173.72***	(11.33)
Wednesday	169.19***	(11.35)
Thursday	150.54***	(11.31)
Friday	84.45***	(11.32)
Saturday	-19.98*	(11.33)
Public holiday	-128.74***	(16.51)
Summer holiday	117.20***	(13.87)
Quarter=2	209.33***	(13.41)
Quarter=3	2.60	(15.98)
Quarter=4	-46.57***	(9.14)
Bike sharing Gdansk	57.66***	(11.85)
Cycling May	*158.55****	*(14.02)*
Constant	368.03***	(20.22)

Counter FE	Y
N	32,850
LL	-253701
R squared	0.70
AIC	507441
BIC	507600

* p<0.1 ** p<0.05 *** p<0.01

Atmospheric conditions, day of the week and period of the year significantly affect bicycle traffic. Cloudiness, wind speed and precipitations negatively affect bike traffic, while more bicycles are observed with increasing temperature. Higher bicycle traffic is observed during weekdays compared to weekends, and less bicycles during public holidays, meaning that bicycles are a way of transport mainly used by commuters, compared to the use for leisure activities. During spring and summer holidays, more bicycles are observed; the implementation of bike sharing system in Gdansk produced a significant increase in bicycle traffic.

Table 11 displays estimates of alternative model specifications, to test the robustness of results. Model (b) takes the natural logarithm of daily bicycle counts; model (c) is the Poisson regression using the level of the count. Besides the average effect produced by the Cycling May intervention, we explore the effect in relation to the number of registered participants in model (d), which can be used as a proxy of the "intensity" of the treatment. The resulting coefficient can be interpreted as the average increase in overall daily number of bicycles per counter generated by each registered participant.

If we consider the average percentage change in model (b), a 18% increase in bicycle traffic is estimated; finally, the ATT estimate from model (c) suggests an increase of 15% in the expected number of bicycles when the intervention is in place. Results of model (d) indicates that for each thousand individuals participating there are on average five bicycle more per day per counter; this results in a total of nearly 4 thousand more bicycles per day in Gdansk (23 counters * 5 bikes* 34 (thousand participants in 2019)= 3,910 daily increase) when the policy is in place, which is consistent with results of other models (e.g. 159 more bikes * 23 counters= 3,657 daily increase).

Table 11 Estimated effect of the Cycling May campaign using a DID approach, alternative specifications

	(b)		(c)		(d)	
Cycling May	*0.18****	*(0.011)*	*0.15****	*(0.010)*		
Participants (th.)					*4.53****	*(0.452)*
Counter FE	Y		Y		Y	
N	32,727		32,850		32,850	
LL	-18640		-1553297		-253715	
R squared	0.88		0.90		0.70	
AIC	37318		3106632		507468	
BIC	37477		3106792		507628	

Standard errors in parentheses; * p<0.1 ** p<0.05 *** p<0.01. Percentage increase (i.e. Exp(beta)-1), Pseudo-R squared and Log pseudo-likelihood reported for model (b). Covariates estimates not shown.

Results are consistent and indicate a significant increase in bicycle traffic in Gdansk attributable to Cycling May campaign. If one relates the estimate of absolute number of additional bicycles (model a) to the average bicycle traffic in Gdansk over the three years of our sample (882 bicycles per day), the estimated impact of the campaign is again 18%, which matches the result obtained from model (b).

Estimates displayed in Table 12 refer to the persistence effect of Cycling May in the days after the initiative ends (model e). The stock variable represents the average daily traffic for each counter in the past 30 days, therefore the interpretation of the stock estimated coefficient is related to persistence: the higher the traffic in the past month, the higher the traffic in the considered day. The interaction of stock, month of June and city of Gdansk reveals the difference between the average persistence over the whole sample, and the specific (additional) persistence in June in Gdansk; the negative and small, but highly significant, coefficient means that in the month after the implementation of the Cycling May campaign there is a decrease in persistence, in other words the increase in bicycle use in May is not sustained after the policy ends. Individuals that used the bicycle in May return to their usual mode of transport and the policy is not effective in changing people behaviour in the long term.

Table 12 DiD estimates - Cycling May persistence

Model (e)	Coefficient	Standard error
Participants (th.)	4.11***	(0.36)
Stock	0.83***	(0.01)
Stock # June # Gdansk	-0.03***	(0.01)
Counter FE	Y	
N	31,950	
LL	-237255	
R squared	0.84	
AIC	474552	
BIC	474727	

Standard errors in parentheses; * p<0.1 ** p<0.05 *** p<0.01.Covariates estimates not shown.

Table 13 reports additional robustness checks. In model (f) the dependent variable is the average daily bicycle count by city, the positive and significant effect of Cycling May is confirmed. Including differential linear trends (model g) does not sensibly change the estimated effect of Cycling May. Lastly, we account for city-specific quarterly differences in bicycle traffic for each year in model (h), and again we find consistent results.

Table 13 DID estimates, robustness checks

	(f)		(g)		(h)	
Quarter1					140.69***	(49.05)
Quarter2					25.61	(39.75)
Quarter3					59.81	(40.20)
Quarter4					281.92***	(38.46)
Quarter5					85.64**	(34.39)
Quarter6					-21.83	(39.81)
Quarter7					97.45**	(40.74)
Quarter8					309.23***	(38.41)
Quarter9					46.99	(34.42)
Quarter10					6.13	(40.02)
Quarter11					8.67	(40.45)
Quarter12					119.68***	(38.52)
Quarter1 # Gdansk					-119.15**	(56.33)
Quarter2 # Gdansk					-57.39	(52.18)
Quarter3 # Gdansk					-44.50	(52.60)
Quarter4 # Gdansk					-101.05**	(45.20)
Quarter5 # Gdansk					-167.34***	(40.17)
Quarter6 # Gdansk					-88.65*	(49.64)
Quarter7 # Gdansk					-126.09**	(50.28)
Quarter8 # Gdansk					-112.93**	(44.09)
Quarter9 # Gdansk					-179.86***	(39.43)
Quarter10 # Gdansk					-89.59*	(48.52)
Quarter11 # Gdansk					-60.96	(50.29)
Quarter12 # Gdansk					58.26	(42.82)
Cycling May	*170.48****	*(24.86)*	*154.41****	*(14.02)*	*146.61****	*(14.60)*
Monthly trend			-3.27***	(0.62)		
Monthly trend # Gdansk			1.14	(0.77)		

Counter FE		Y	Y
City FE	Y		
N	2,190	32,850	32,850
LL	-14826	-253676	-253622
R squared	0.88	0.70	0.70
AIC	29688	507395	507323
BIC	29791	507571	507650

Quarters are three month-periods (Jan-Mar, Apr-June, Jul-Sep, Oct-Dec), except Quarter1= September 2016; Quarter13=July and August 2019. Quarter13 is the reference period.
Standard errors in parentheses. * p<0.1 ** p<0.05 *** p<0.01. Covariates estimates not shown.

3.5 Selection unobservables: synthetic control methods

When pre-policy data cover multiple periods and multiple non-treated groups (e.g. regions), the synthetic control method (SCM) is a popular option (Abadie et al., 2015) . Consider a situation where only one region is treated, and there are *n* non-treated regions. The principle is relatively straightforward, instead of using the *n* controls separately, they are artificially combined into a single control group as a weighted average. The weights are obtained through an optimization algorithm which minimizes – in each time period before the policy – the distance between the outcomes and the observed covariates measured in the target group and those obtained as the weighted average of the *n* values measured in the multiple control groups. In other words, the SCM allows not only to ensure the common trend between the treated region and the artificial control group, but also balances the covariates. Then, the weights can be applied in the post-policy period to obtain the counterfactual outcomes.

This method "extends the traditional linear panel data (difference-in-differences) framework, allowing that the effects of unobserved variables on the outcome vary with time"(Abadie et al., 2010)

Data requirements:

- Balanced panel dataset including
    - a pre-intervention period and a post-intervention period
    - one unit (e.g. a region) exposed to the intervention and multiple units not exposed to the intervention (donor pool). If necessary data can be first aggregated (e.g. from the household level to the regional level).
- A vector of observed covariates which are reasonable predictors of the outcome

Application 7. Estimating the effect of the Catalonia soda tax on CPIs, allowing for differential time trends under a Synthetic Control approach

Dataset: CPI.dta

Outcome: CPIs for non-alcoholic drinks

Treated unit/s: Catalonia

Control unit/s: remaining 18 autonomous regions

Strategy: a synthetic control region is constructed as a data-driven weighted average of the non-treated regions. Pre-treatment CPIs and other time invariant observed covariates are used.

Table 14 reports the average of CPIs predictors for Catalonia, for the synthetic control region and for the 18 real control regions. This gives an idea of the similarities (or lack thereof) between Catalonia and the synthetic control region, in terms of pre-tax predictors of post-tax outcomes.

Table 14 CPIs predictors: average values for Catalonia, average values for the synthetic control region and average values across non-taxed regions (pre-tax).

	Catalonia		
	Real	Synthetic	Average of 18 control regions (donors)
Household size	2.562	2.587	2.654
One-person household	0.183	0.188	0.190
One parent with children less than 16 y.o.	0.025	0.022	0.024
One parent with children older than 16	0.054	0.067	0.074
Couple without children	0.281	0.280	0.249
Couple with children less than 16 y.o.	0.219	0.212	0.195
Couple with children older than 16	0.172	0.172	0.191
Other household types = o,	0.065	0.062	0.077
Age of the household reference person	54.481	55.229	55.128
Education level 1	0.141	0.155	0.205
Education level 2	0.300	0.298	0.305
Education level 3 = o,	0.220	0.209	0.185
Education level 4	0.339	0.341	0.305
Pensioner-only household	0.282	0.295	0.278
Food - Average purchased quantity	2403.713	2464.832	2502.844
Non-alcoholic drinks - Average purchased quantity	455.804	432.541	457.416
Alcoholic drinks -Average purchased quantity	165.345	173.231	160.456
Non-alcoholic softdrinks -Average purchased quantity	162.372	174.369	181.109

Table 15 shows the relative contribution of each control region to the counterfactual synthetic region. “Because the weights can be restricted to be positive and sum to one, the synthetic control method provides a safeguard against extrapolation”(Abadie et al., 2010)

Table 15 Region weights in the synthetic Catalonia

Control region	Weight	Control region	Weight
Andalucia	0	Extremadura	0
Aragón	0.375	Galicia	0
Asturias	0	Madrid	0.149
Balearic	0.058	Murcia	0
Canary	0	Navarra	0.197
Cantabria	0	Basque	0
Castile_leon	0	Rioja	0
Castile_mancha	0	Ceuta	0
Valencia	0.173	Melilla	0.05

According to Figure 8, CPI for non-alcoholic drinks in the synthetic Catalonia very closely track the trend of the same CPI in Catalonia for the pre-tax period. Given the high degree of balance on all CPIs predictors shown in Table 15, this suggests that the synthetic Catalonia provides a good approximation to the CPIs that would have been registered in Catalonia after May 2017 in the absence of the tax. Immediately after May 2017, the two lines begin to diverge clearly, which suggests a large negative effect of the tax on CPIs.

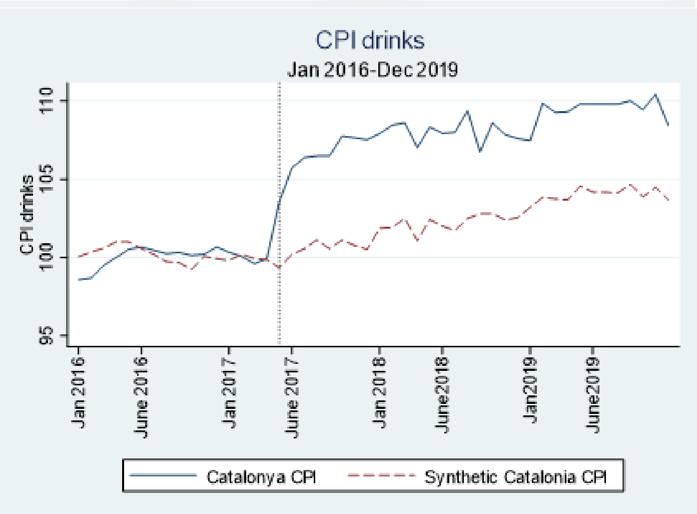


Figure 8 Trends in CIPs: Catalonia vs synthetic Catalonia

3.6 Selection on unobservables, randomization with imperfect compliance: the instrumental variable (IV) approach

$\mathrm{D}(\mathrm{X},\mathrm{U},\mathrm{Z}) = \mathrm{D}(\mathrm{U},\mathrm{Z})$ where U is correlated to Y^0.

In some cases, even if the selection to the treatment is randomized (D depends on a random binary variable Z), there might be imperfect compliance to the treatment assignment. As a result, the actual treatment state does not coincide with the assigned treatment state. If the compliance mechanism is not known it is not possible to estimate the ATT. However, it is possible to estimate:

- the Intention to treat (ITT), which is the causal effect of being assigned to the treatment (not necessarily being treated). The estimation requires comparing the two groups indexed by Z: $\{Y|Z{=}1\} - \{Y|Z{=}0\}$. No selection bias arises since Z is randomly determined.
- The Local Average Treatment Effect (LATE): the causal effect of the treatment on compliers. This is obtained by using Z as an instrumental variable for D

Provided that one or more "good" instruments are available, IV estimators of the ATT allow to control for selection effects driven by both observables and unobservables. Let us consider a general outcome equation, where

$$y_i = \alpha + \beta D_i + \gamma \boldsymbol{x_U} + \delta \boldsymbol{x_U} + \varepsilon_i \tag{21}$$

where y_i is the outcome for the *i-th* unit, D_i is a binary indicator (the *policy dummy*) which is equal to 1 when the *i-th* unit is exposed to the policy and 0 otherwise, $\boldsymbol{x_O}$ is a vector of observed variables and $\boldsymbol{x_U}$ a set of unobserved variables (equation 21 differs from equation 7 because of the introduction of the latter).

In absence of information on $\boldsymbol{x_U}$, we face the econometric textbook problem of omitted variables, so that all coefficient estimates are biased and inconsistent. Under an economics viewpoint, a parallel interpretation is that the selection variable D_i is endogenous, as the probability of being exposed to the policy depends on the outcome level. For example, schools located in high income and education areas where fruit consumption is high, are more likely to participate in school fruit schemes.

Provided we have one or more adequate instruments $\boldsymbol{w}$ to instrument D_i we can control for the selection bias and obtain consistent ATT estimates, at the cost of giving up some efficiency. Statistical packages routinely provide IV-2SLS estimators where the first stage regression is again a binary dependent variable model, a probit or a logit. Note that the structural policy model 21 still accounts for unbalances in observables x_O , which enter directly the model as they are expected to influence the outcome. Instead, instruments should be variables that we would not use as direct explanatory variables for the outcome, and should be exogenous. If we have access to such type of variables, the first stage binary regression would be the same used to estimate propensity scores, with x_O as explanatory variables, plus the instruments $\boldsymbol{w}$ which do not belong to x_O and do not enter equation 21.

Since IV encompasses PSM and accounts for selection on unobservables, why don't researchers just rely on IV estimation? The problem is likely to be a familiar one for the experienced reader. First, we struggle to find reasonable instruments in the dataset. Second, we struggle to convince reviewers that our instrument choice is a good one. Unfortunately, there is no definitive test on the validity of instruments that can convince all actors in the publication process. The issue is a Catch-22 one. In order to show that an instrument is exogenous, it must be independent from the residuals of the structural (second stage) equation. However, this test is theoretically impossible, as we only obtain unbiased estimates of the residuals if we have an exogenous instrument. The empirical solution is to use several instruments, leave one out, estimate the structural equation residuals through the other instruments, then check the correlation between the excluded instrument and the estimated residuals. One can then repeat the procedure leaving out a different instrument each time. While such a strategy may provide some support to the instrument validity claims, it is an empirical one, and it is still grounded on the assumption that the included instrument are exogenous and the residual estimates are unbiased. If many of our instruments are endogenous, the procedure is useless. Thus, we still need to be convinced and convince others that the instruments make sense under an economic perspective.

The other interesting element is the trade-off between consistency and efficiency. If the instrument are reasonable, exogenous, and obviously significant in the first stage equation, then we can place some trust in the consistency of the ATT estimate in the second stage equation. However, the ATT will have a larger standard error, as we rely on predictions of the Di variable in the second stage, a sort of propensity scores augmented by the instruments. How much larger the standard error depends again on the goodness-of-fit of the first stage probit or logit equation. This time, however, a poor fit does not lead to systematic biases, it just inflates the standard errors, and with large data-sets this is not usually a problem.

A list of instruments used in the food policy literature is beyond the scopes of this article, although it would be an interesting reading. For example, Hofferth & Curtin, 2005 investigate the effect of school lunch programs on the BMI of students. Participation to the lunch programs is voluntary for schools, and students need to have specific characteristics to be eligible for a free meal. These policy elements are clearly a source of endogenous selection. Public school attendance is used as an instrument, as it does not affect BMI directly, but it is strongly associated with the school program participation, since public schools are more likely to be part of lunch programs.

An alternative strategy resting on the use of instruments is the control function approach. This approach involves a first stage to model the exposure to the program, and a second stage where the individual probability of exposure is included as an additional variable on the right-hand side of the outcome model, to correct for the selection bias. The Heckman two-step estimator is the most widely used control function approach. For example, Butler & Raymond, 1996 explore the impact of household participation in US Food Stamp program on nutrient intakes of the elderly, using a variety of instruments, including household assets and distance to a food stamp office.

3.7 Strategies based on structural models

An alternative approach is needed in situations where there is no natural counterfactual, for instance when a policy potentially acts on the whole population, as in a nationwide a public information campaign. As information policies may be expected to generate behavioural effects beyond the mere change of the average outcome, an option is to generate model-based counterfactual estimates. This approach is especially interesting when the behaviour of interest is well captured by a consolidated economic specification, and it is conveniently applicable when the pre-policy and post-policy data come from different (repeated) cross-sectional samples from the same population. One may then express the outcome as the function of its determinants in each period:

$$y_i^0 = f^0(\mathbf{x}_{i0}^0) + \varepsilon_i^0 \qquad 22$$

and

$$y_i^1 = f^1(\mathbf{x}_{i0}^1) + \varepsilon_i^1 \qquad 23$$

The functions f^0 and f^1 have the same structural specification, but are characterized by different parameters. For example, f might be a demand function and the parameters represent price and income elasticities. As implied by the Lucas critique, a policy is likely to go beyond changing the average level of consumption, and also lead to a change in elasticities, hence the change from f^0 to f^1.

If the policy has no direct impact on the covariates $\mathbf{x}_{i0}$, then he two set of estimates allow to evaluate the counterfactual outcome, which is estimated as $\tilde{y}_i^1 = f^0(\mathbf{x}_{i0}^1)$. In in our example this is the level of consumption that would have been observed in period 1 had the population maintained the preference structure of period 0. The ATT is $f^1(\mathbf{x}_{i0}^1) - f^0(\mathbf{x}_{i0}^1)$. The approach can be modified to include constraints on behavioral parameters, for example one might require that some of them remain constant between the two time periods. Also, if there are variables in $\mathbf{x}_{i0}^1$ that are significantly affected by the policy, and it is possible to disentangle such effect (e.g. an estimate of the change in public advertising expenditure, or of the price change associated with a tax), one might estimate the counterfactual through $f^0(\hat{\mathbf{x}}_{i0}^1)$ where the relevant variables in $\mathbf{x}_{i0}^1$ are purged from the policy effect.

When data are organized as panels or relatively long time series, alternative approaches based on structural models may rely on switching and time-varying parameter regressions, intervention or event study analyses. All of these models allow one or more parameters to change in response to the policy. The most basic formulation aims at estimating a sharp step (i.e. an intercept shift as in event studies) at the time of the policy implementation. When data allow to do so, any parameter in the structural model can potentially change and evolve, either with a pre-determined shape (as in intervention analysis or switching regression) or through random shocks (as in time-varying parameters models).

An example of nutrition policy evaluation where the counterfactual is based on a structural model is provided in (Capacci & Mazzocchi, 2011), who explore the effects of the 5-a-day information campaign in the UK through a demand system. (Attanasio et al., 2012) exploit randomization in the Mexican program Progresa to discuss how structural models can improve program evaluations even in cases where evidence from experiments is available. (Kim et al., 2001) exploit a switching regression model to estimate the effect of the Nutrition Labelling and Education Act on diet quality in the US.

3.8 Other methods and extensions

When estimating real-world policy impacts, it is important to consider that the actual impact – or treatment effect (TE) – of the policy may be very heterogeneous across exposed subjects due to various reasons, and average estimates (ATE) may thus be unsatisfactory. If subjects are exposed to the policy, but do not comply with the intervention, ATE estimates become problematic, as non-compliers are likely to systematically differ from both compliers and control subjects (i.e. reasons for compliance are correlated with TE). Consequently, two different TEs can be estimated: (1) considering all those exposed regardless of their compliance, which returns the average intention-to-treat (ITT) effect; and (2) considering treated subjects only, while accounting for the additional selection bias, which returns the local average treatment effect (LATE). When non-compliance is an issue, the LATE can be obtained through an IV estimator (Imbens and Wooldridge, 2009). Furthermore, TEs may be heterogeneous between subjects due to the nature of the intervention (as in e.g. personalized nutrition or physical activity programs) since its effectiveness primarily depends on subject characteristics. Recently, there is a growing interest in methods (mostly based on machine learning) that capture this heterogeneity of policy impact across subpopulations, by letting the ATE depend on sample covariates (CATE), Wendling et al 2018 . Quantile DID models (Chakrabarti et al 2018) and LASSO estimators may be of use for the quantification of heterogeneous treatment effects (Belloni et al 2017).

Applications and future directions

Within the PEN project, ongoing work is exploring the variable impact of the tax on the unit values paid by consumer in Catalunya as a response to the SSB tax. More specifically, a RIF-DID (Recentered Influence Function-DID) method is used (details will be provided in a forthcoming paper by Capacci, Calia, Ferrante).

4. List of References

Abadie, A., Diamond, A., & Hainmueller, A. J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, *59*(2), 495–510. https://doi.org/10.1111/ajps.12116

Abebaw, D., Fentie, Y., & Kassa, B. (2010). The impact of a food security program on household food consumption in Northwestern Ethiopia: A matching estimator approach. *Food Policy*, *35*(4), 286–293. https://doi.org/10.1016/J.FOODPOL.2010.01.002

Aguilar, A., Gutierrez, E., Seira, E., & Itam, M. (2021). The effectiveness of sin food taxes: Evidence from Mexico. *Journal of Health Economics*, *77*, 102455. https://doi.org/10.1016/j.jhealeco.2021.102455

Angrist, J. D., & Pischke, J. S. (2014). Mastering 'metrics: The path from cause to effect. In *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press. https://doi.org/10.1093/erae/jbv011

Attanasio, O. P., Meghir, C., & Santiago, A. (2012). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA. *Review of Economic Studies*, *79*(1), 37–66. https://doi.org/10.1093/restud/rdr015

Beatty, T. K. M. (2008). Expenditure dispersion and dietary quality: Evidence from Canada. *Health Economics*, *17*(9), 1001–1014. https://doi.org/10.1002/hec.1393

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C. Program Evaluation and Causal Inference With High-Dimensional Data. Econometrica 2017;85:233-298Braha, K., Cupák, A., Pokrivčák, J., Qineti, A., & Rizov, M. (2017). Economic analysis of the link between diet quality and health: Evidence from Kosovo. *Economics & Human Biology*, *27*, 261–274. https://doi.org/10.1016/J.EHB.2017.08.003

Butler, J. S., & Raymond, J. E. (1996). THE EFFECT OF THE FOOD STAMP PROGRAM ON NUTRIENT INTAKE. *Economic Inquiry*, *34*(4), 781–798. https://doi.org/10.1111/J.1465-7295.1996.TB01410.X

Caliendo, M., & Kopeinig, S. (2008). SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/J.1467-6419.2007.00527.X

Capacci, S., & Mazzocchi, M. (2011). Five-a-day, a price to pay: An evaluation of the UK program impact accounting for market forces. *Journal of Health Economics*, *30*(1), 87–98. https://doi.org/10.1016/j.jhealeco.2010.10.006

Capacci, S., Mazzocchi, M., & Shankar, B. (2018). Breaking Habits: The Effect of the French Vending Machine Ban on School Snacking and Sugar Intakes. *Journal of Policy Analysis and Management*, *37*(1), 88–111. https://doi.org/10.1002/PAM.22032

Chakrabarti S, Kishore A, Roy D. Effectiveness of Food Subsidies in Raising Healthy Food Consumption: Public Distribution of Pulses in India. American Journal of Agricultural Economics 2018;100:1427-1449.

Clark, M. A., & Fox, M. K. (2009). Nutritional Quality of the Diets of US Public School Children and the Role of the School Meal Programs. *Journal of the American Dietetic Association*, *109*(2), S44–S56. https://doi.org/10.1016/J.JADA.2008.10.060

Colchero, M. A., Salgado, J. C., Unar-Munguía, M., Hernández-Ávila, M., & Rivera-Dommarco, J. A. (2015). Price elasticity of the demand for sugar sweetened beverages and soft drinks in Mexico. *Economics & Human Biology*, *19*, 129–137. https://doi.org/10.1016/J.EHB.2015.08.007

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34*, 271–310. https://doi.org/10.1111/j.0081-1750.2004.00154.x

Frölich, M., & Huber, M. (2019). Including Covariates in the Regression Discontinuity Design. *Journal of Business and Economic Statistics*, *37*(4), 736–748. https://doi.org/10.1080/07350015.2017.1421544/SUPPL_FILE/UBES_A_1421544_SM7297.PDF

Hausman, C., & Rapson, D. S. (2018). Regression Discontinuity in Time: Considerations for Empirical Applications. *Https://Doi.Org/10.1146/Annurev-Resource-121517-033306*, *10*, 533–552. https://doi.org/10.1146/ANNUREV-RESOURCE-121517-033306

Hofferth, S. L., & Curtin, S. (2005). Poverty, food programs, and childhood obesity. *Journal of Policy Analysis and Management*, *24*(4), 703–726. https://doi.org/10.1002/PAM.20134

Imbens GW, Wooldridge JM. Recent Developments in the Econometrics of Program Evaluation. J Econ Lit 2009;47:5-86

Kim, S. Y., Nayga, R. M., & Capps, O. (2001). Food label use, self-selectivity, and diet quality. *Journal of Consumer Affairs*, *35*(2), 346–363. https://doi.org/10.1111/J.1745-6606.2001.TB00118.X/FORMAT/PDF

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/PAN.2019.11

Kleven, H. J. (2016). Bunching. In *Annual Review of Economics* (Vol. 8, pp. 435–464). Annual Reviews. https://doi.org/10.1146/annurev-economics-080315-015234

Kraus, S., & Koch, N. (2021). Provisional COVID-19 infrastructure induces large, rapid increases in cycling. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15). https://doi.org/10.1073/PNAS.2024399118/-/DCSUPPLEMENTAL

MacPherson, C., & Sterck, O. (2021). Empowering refugees through cash and agriculture: A regression discontinuity design. *Journal of Development Economics*, *149*, 102614. https://doi.org/10.1016/J.JDEVECO.2020.102614

Rahkovsky, I., & Gregory, C. A. (2013). Food prices and blood cholesterol. *Economics & Human Biology*, *11*(1), 95–107. https://doi.org/10.1016/J.EHB.2012.01.004

Schanzenbach, D. W. (2009). Do School Lunches Contribute to Childhood Obesity? *Journal of Human Resources*, *44*(3), 684–709. https://doi.org/10.3368/JHR.44.3.684

Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Stat Med 2018;37:3309-3324.

You, J. (2013). The role of microcredit in older children's nutrition: Quasi-experimental evidence from rural China. *Food Policy*, *43*, 167–179. https://doi.org/10.1016/J.FOODPOL.2013.09.005