Dottorato in Economics - Ciclo XXXII

Università degli Studi di Milano e Università degli Studi di Pavia

ESSAYS ON POLICY EVALUATION

Candidato: Fabio I. Martinenghi

Relatori: Prof. Michele De Nadai

Prof. Giovanni Pica

Anno Accademico 2018/2019

Essays in Policy Evaluation

Fabio I. Martinenghi

1 Introduction

This thesis comprises two rather different chapters. The first chapter explores the impact of increasing the exit costs of cohabitation on the stability of relationships. The second chapter explores the impact that an established environmental beach award has on the tourism sector and the balance sheet of a municipality. Notwithstanding these thematic differences, the two chapters are connected by the way in which the research therein was conducted. This particular approach to research is known as the "credibility revolution". It was championed by Angrist and Pischke in Angrist and Pischke (2008) and Angrist and Pischke (2010). The approach is concerned with causality above all other issues and aims at minimising the assumptions made when estimating an object of interest; be it theoretical or technical. In order to construct well-founded economic models of behaviour based on the "credibility revolution", this sort of humble policy evaluation work needs to be abundant. Without gathering sufficient evidence about the causal mechanisms that inform the behaviour of individuals; there can be no strong foundations upon which to build sound economic models. This is only likely to lead to misleading analyses and policy failures.

2 Economists as Plumbers?

The "credibility revolution" is also connected with what Nobel Laureate Esther Duflo characterised as the "plumber economist" (Duflo, 2017). Both approaches are concerned with estimating causal relationships. The concept of the economist as a plumber goes one step further than the "credibility revolution" when it comes to evaluating policies. The former focuses on taking all of the policy details into account, ranging from the complex legal aspects to the practical implementation of it.

Despite the work an advocacy of scholars as Angrist and Duflo, (Cherrier, 2016) argues that most of this empirical turn in economics is more due to an increase in applied theory than empirics. Cherrier argues that in the past decades, we have witnessed the demise of pure theory in favour of theory *with* data, rather than a real change of attitude towards *empirical* work. While the credibility revolution has made the identification of causality the goal of any applied work, theoretical work retains a superior prestige independently of how solid it is, as I show below.

3 Theory and Data

In an ideal empirical science, inductive inference occupies a crucial role, especially when theory needs to be generated or improved. Guala et al. (2005) identifies two phases in scientific research, as visualised in Figure 1. The first phase involves identifying a phenomenon in the data, which are by their own nature noisy. Using an example relevant to my research, do cohabiting couples respond to changes in the exit costs of cohabitation? If so, how?

This first phase of scientific research involves a measurement task. It is the task of measuring, for instance, how the probability of separation changes for a cohabiting couple when the legal costs of separating increase. Once this phenomenon has been measured nationally and internationally, researchers can progress to phase two.

In phase two, the *why* question is asked. Going back to my example, if couples are found to consistently respond in a certain way to changes in the exit costs of cohabitation, we can ask ourselves why. It is at this point that theory comes into play, providing a systematic explanation of the phenomenon and connecting it with other phenomena. With respect to the above example, this would be a theory of how people choose their partner, how they choose between cohabitation and marriage and how they choose whether to terminate a partnership.

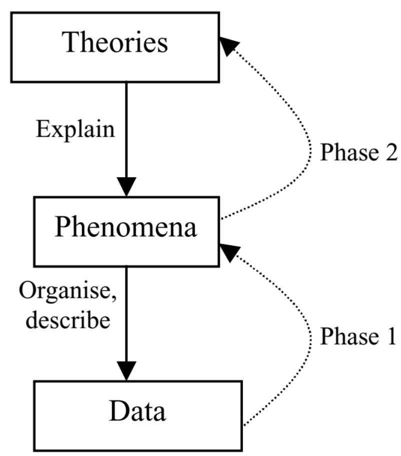


Figure 1. Relationship between data, phenomena and theories, Guala et al. (2005)

However, contemporary economics is still true to its *ultra-deductivist* roots (Hutchison, 1998), at least since the seminal works of (Mill, 1994) and then (Robbins, 1994). The standard approach to the empirical enterprise is therefore through a mathematical model. In most well-published applied papers, first, a model is presented and, second, some empirically observable implications of the model are tested using data. This very structure elicits the preference of economists. Theory comes first, then comes the data. However, empirical work such as policy evaluation studies hardly start from strong theoretical assumptions, embedding them in a model. These kind of studies are measurement exercises, and of the most relevant kind, given the closeness between economics and policy-making. For this reason "imposing" (via publication incentives) this deductivist structure to purely empirical papers does not reflect how the research was conducted in practice.

If a researcher's approach is strictly empirical, as defined above, any modelling exercise used to rationalise the findings should come after said findings. This kind of research is also called "exploratory" (Jupp, 2006). Despite their importance as the foundation of any evidence-based theory, empirical studies are not considered as prestigious as applied ones, unless they uncover a particularly interesting phenomenon.

These two types of studies are different and both vital for the cumulation of economic knowledge. To see this clearly, it is useful to go back to (Guala et al., 2005) and follow him in the use of Alvin Roth's taxonomy of experiments (Roth, 1988), which can be applied by analogy to observational studies too. Applied work belongs to the *Speaking to theorists* kind of research, which aims at testing hypothesis derived from theoretical models. Empirical work belongs to the *Searching for facts* research, which explores new phenomena[1], with particular interest in those phenomena that cannot be reconciled with current theories. Roth calls the final category *Whispering in the ears of princes*. This category includes all research aiming at aiding policy making and is hardly considerable in isolation. Indeed, most economic papers include a section or paragraph dedicated to policy implications.

We live in a period of reformation in economics, with pressure mounting to improve the field over a range of elements, all functional to the production of better economic science. Open-source software — namely R and Python — is becoming more common in the field. Top journals are starting to require researchers to send, together with their papers, data and code for publication. The creation of the journal *Series of Unsurprising Results in Economics* (SURE) is a first response to the dissatisfaction with the field's bias against publishing null findings. Let us add one more element to this Economics reform agenda: recognising the crucial importance of the most empirical work, which humbly provides a solid foundation to the whole field. Only by doing this, can we keep improving our models of human behaviour and discard those inconsistent with the evidence.

[1]Note that to create consensus in the scientific community around the existence of a new phenomenon, several empirical studies replicating it need to be conducted.

References

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

Cherrier, B. (2016). Is there really an empirical turn in economics? [/www.ineteconomics.org; Visited on 1st-November-2019].

Duflo, E. (2017). Richard t. ely lecture: The economist as plumber. *American Economic Review*, 107(5):1–26.

Guala, F. et al. (2005). *The methodology of experimental economics*. Cambridge University Press.

Hutchison, T. (1998). Ultra-deductivism from nassau senior to lionel robbins and daniel hausman. *Journal of Economic Methodology*, 5(1):43–91.

Jupp, V. (2006). *The SAGE Dictionary of Social Research Methods*. SAGE Publications, London.

Mill, J. S. (1994). On the definition and method of political economy. In Hausman, D., editor, *The philosophy of economics: An anthology*, chapter 1, pages 52–68. Cambridge University Press, Cambridge.

Robbins, L. (1994). The nature and significance of economic sciencey. In Hausman, D., editor, *The philosophy of economics: An anthology*, chapter 3, pages 73–99. Cambridge University Press, Cambridge.

Roth, A. E. (1988). Laboratory Experimentation in Economics: A Methodological Overview. *The Economic Journal*, 98(393):974–1031.

Chapter I
De facto marriage: when ending a cohabitation costs as much as a divorce.

March 31, 2020

1 Introduction

Despite the seminal work by Becker (1973) started a prolific research agenda on the economic aspects of marriage, the validity of applying methodologies conceived to study economic behaviour to family formation remains problematic. Already in the Nineteenth century, John Stuart Mill warned that while economics is solely interested in man's desire for wealth and its efficiency in obtaining it, there are situations in which such desire is in competition with other principles of human nature. Such non-economic motives, Mill maintained, act as *confounding causes*, potentially invalidating the deductive (*à priori*) results of economic models (Mill, 1994), as they do not take them into account. Mill's recommendations are still valid today. In particular, they justify the importance of empirically supporting or falsifying the theories that social scientists have constructed on how couples choose between informal cohabitation and marriage (see **???**). Indeed, the causal mechanisms at the heart of these theoretical models describing how and why we partner need further empirical testing (Matouschek and Rasul, 2008, being the only paper to my best knowledge addressing this issue). The scarcity of quasi-experimental evidence has left economists struggling to identify those causal links. This is critical, for instance, to understand what kind of legislation can promote the formation of stable unions.

This paper contributes to fill that gap in the literature, using a natural experiment to study how increasing the exit costs of cohabitation affects the formation and stability of couples. This is particularly relevant as the past decades have seen both household formation and dissolution changing dramatically in developed countries. As the demographic literature shows (see Perelli-Harris et al., 2017), while cohabitation has emerged as a new and trending way to live a romantic relationship, marriage has become less frequent and less stable. I look at the effects of making the exit costs of cohabitation as high as divorce and how this impacts new unions[1] and existing ones. Specifically, I exploit the Family Law Amendment Act,

[1] a couple that either cohabits or is married or both at different times

introduced in Australia in 2008, as an exogenous shock to the cost of exiting cohabitation.

The Family Law Amendment Act 2008 (Parliament of Australia, 2008) became active in 2009 and defined couples "living together on a genuine domestic basis"—i.e. acting as a married couple—as *de facto relationships*. Under the Family Law Amendment Act 2008, the termination of a de facto relationship carries the same consequences as getting divorced, for example giving to the spouses the right to seek a property settlement and spousal maintenance. I hence exploit the discontinuity in time produced by the reform to identify its effects on the stability of new and existing couples. I find that when terminating a cohabitation becomes as costly as getting divorced, (i) new unions are more stable (ii) existing cohabitors affected by the reform in their third year are more likely to split, while (iii) the probability of starting a cohabitation and the duration of premarital cohabitation do not change.

Previous research on the effects of changing the cost of ending a union has focused on marriage, particularly on reforms moving from mutual consent divorce to unilateral divorce regimes (Friedberg, 1998; Wolfers, 2006; Lee and Solon, 2011). In this paper, I add to that literature by focusing on changes in exit costs from cohabitation. This is particularly relevant as getting married after a period of cohabitation is increasingly becoming the norm. In countries like Norway, Spain (see Perelli-Harris et al., 2017) and Australia (Hewitt et al., 2005) this is already the case.

Furthermore, changing the cost of terminating a union affects new and existing couples in different ways. This has been shown analytically by Matouschek and Rasul (2008) in reference to divorce law reforms. In particular, on the one hand, new regimes will incentivise the formation of couples of a certain quality and disincentivise others, hence the composition of the couples starting a union before and after the reform will be different. They call this "selection effect". On the other hand, these policies change the incentives of already existing couples. They call this "incentive effect". While in their empirical analysis they are able to identify the incentive effect, they do not fully identify the selection effect.

I test the hypothesis that a higher expected exit cost from cohabitation will deter low quality matches (couples) from entering cohabitation. This is a *selection effect* hypothesis. For these low quality couples, the expected probability of separation is high, hence their net benefit from cohabiting will decrease more given an increase in the exit costs. This implies that the average match quality is going to increase after the reform, observable as a lower probability to separate for new couples. I am able to separate the selection channel by comparing couples which started in the three years before the reform with those which started within three years from the reform and comparing them only for those periods in their relationships in which they were both under the reform. In particular, I can identify the selection effect by comparing couples that live under the same new legal regime but which started under different regimes.

On the incentive effect side, I improve on the identification the incentive channel by comparing couples that were in their j^{th} year of cohabitation just before the reform with the cohabiting couples that were in their j^{th} year just after. However, the timing and sign of the incentive effect is controversial ex-ante. Specifically, under the Family Law Amendment Act

2008, couples who decide to move in together are not automatically de facto relationships. They become de facto later on, but they do not know exactly when. Hence it is more challenging to make predictions on the incentive effect of the policy. Indeed, the reform applies only to cohabiting couples "living together under a genuine domestic basis". This concept is not well-defined, but it is more likely to apply the more the cohabitation lasts. Due to this difficulty in predicting the incentive effect, I take a more data-driven approach and estimate it using a flexible specification. I find that only cohabitors affected by the reform while in their third year of the relationship are more likely to separate that year. This might be a threshold emerging spontaneously, given the lack of a formal one. In other words, couples affected by the reform in their third year might see it as their last possibility to break up before being considered as married, causing lower quality couples to break up.

Other findings are more difficult to rationalise, particularly without departing from neo-classical assumptions, i.e. without introducing some behavioural assumptions. First, the duration of premarital cohabitation is not affected by the reform. On the contrary, once marriage and cohabitation are equalled and cohabitation loses its flexibility, we would expect premarital cohabitation time to significantly reduce for new couples. Secondly, the number of new cohabitations remains stable after the reform, while all standard models predict that it should change with the change in exit costs (Matouschek and Rasul, 2008).

This paper contributes to the literature in two ways. First, it is one of the first papers focusing on cohabitation regulations in a causal ways (Chiappori et al., 2017, is the only other one) and the first to look at exit costs from a cohabitation point of view. Secondly, it fully disentangles the selection and incentive channels, the two channels through which any family law reform affects unions.

The remainder of the paper is organised as follows. In Section 2 I give some definitions and present the reform of interest. I then present the data and summary statistics in Section 3. In Sections 4 and 5, I introduce the identification strategy for each of the two channels and then present the empirical specifications and the associated results and robustness checks. In Section 6, I study the transition from cohabitation to marriage and in Section 7 I estimate the changes in the probability of starting a new cohabitation. Section 8 concludes by summarising the findings and deriving some policy implications.

2 Background

2.1 Definitions

For the sake of clarity, it is important to define some concepts used throughout this paper. These are not new definitions. A *cohabitation* is defined as a romantic relationship in which the partners reside in the same dwelling. A *de facto relationship*, or *de facto*, is a cohabitation to which a legal recognition has been granted. Etymologically, the terms *de facto spouses* or *de facto marriage* indicate situations in which a couple is *in practice* living as if it were married. In the context of current Australian Commonwealth law, it indicates two individuals who “have a relationship as a couple living together on a genuine domestic basis” (Parliament of Australia, 2008).

Furthermore, the term *union* is used as a term including both cohabitation and marriage. This can be helpful, since in practice there is wide spectrum of long-term relationships, different in the meaning the couple gives to it. This continuum of union-types, can sometimes make categorisations arbitrary. For instance, a relationship that started as a cohabitation and then became a marriage is considered as a single union.

The term *separation* is both used in its legal meaning, as the moment in which a couple decides their marriage is over, and for indicating the termination of a cohabitation.

Lastly, the term ‘period’ is used when referring to the years of duration of a union (first, second, third, etc.), where confusion with calendar years might arise.

2.2 The 2008 Family Law Amendment

In 1984, the Parliament of New South Wales passed the De Facto Relationship Act (NSW Government, 1984), giving *de facto* partners virtually the same rights as married couples. Until recently, the Parliament of Australia could rule over married but not over *de facto* couples. It was only in November 2008 that the Constitution was modified to include within the power of the federal government the jurisdiction over *de facto* relationships matters. This allowed an amendment to the 1975 Family Law Act to be passed (Parliament of Australia, 2008) which effectively extended the De Facto Relationship Act NSW Government (1984) to the rest of Australia. This policy is interesting because (i) it grants de facto couples the same rights and duties as married couples and (ii) it does so through a loosely defined automatic mechanism. By this I mean that there is not a clear formal rule defining what constitutes a de facto relationship. Indeed, the Family Law Amendment (Parliament of Australia, 2008) establishes that the circumstances defining a de facto “may include any or all of the following: (a) the duration of the relationship; (b) the nature and extent of their common residence; (c) whether a sexual relationship exists [...]” and other six criteria, whilst specifying that “no particular finding in relation to any circumstance is to be regarded as necessary in deciding whether the persons have a de facto relationship.” I argue that this reform causes a behavioural change in those individuals starting a long-term relationship after its introduction.

3 Data

Household and individual data are taken from the *Household, Income and Labour Dynamics in Australia* (HILDA) Survey, which follows the lives of more than 17,000 Australians once a year, starting in 2001. Based on an initial sample of 7,682 households, it follows their lives over the generations, as the children of the initial families create new households. The sample was further extended in 2011, adding 2,153 responding households to counterbalance attrition. The Melbourne Institute designed and manages the study, which records information on a wide range of variables related to the economic life, the psychological well-being and the family dynamics of its participants.

3.1 New South Wales

As detailed in Section 2.2 New South Wales (NSW) was already under a legislation equivalent to the Family Law Amendment Act 2008, since 1984. This might make NSW seem suitable as a *control* state in a difference-in-differences setting. However, it is unclear whether the wave of media coverage of the 2008 reform constitutes for NSW a second treatment (after the 1984 one). If that were the case, the policy impact would be downward biased if not cancelled out. For this reason, I drop unions *started* in NSW when evaluating the selection effect and I drop unions *ongoing* in NSW when evaluating the incentive effect.

3.2 Descriptive Statistics

Table 1 summarises the composition of the unions in the selected sample. The sample includes only unions which formed outside of NSW and which began after the year 2000. It contains data on 3,963 unions of 5,350 individuals over 17 waves, between calendar years 2001-2017. The data is at the individual level, so that if both partners in a union are in the sample then the union is reported twice. The top part of the table reports the observed mean duration of unions, marriages, premarital cohabitation and cohabitation without marriage. Durations are short on average in part due to the right-censoring of the data. Furthermore, as shown in the the second part of the table, cohabitation makes up 60% of the unions. Their frequency, combined with their short average duration of 3 years, lowers the average duration of unions. In the *Partners' Characteristics* part of the table, the covariates used in the analysis are introduced. For categorical variables, the mode is reported rather than the mean. Birth cohort is an ordered categorical variable, grouping all the individuals born in the same decade. For instance, it is equal to 1970 if individual i was born between 1970-1979. The other covariates are only used in Section 4.6 to check if the results are robust to the introduction of divorce predictors common in the literature (see Hewitt et al., 2005). *Remoteness of area* takes integer values from 1 to 5, depending on how remote the area of domicile is. *Relative disadvantage* is a variable derived from one of the Australian Bureau of Statistics' socio-economic indicators for areas from the 2001 census (ABS, 2011). It takes integer values from 1 to 10, each representing which decile an area located in the index of socio-economic disadvantage is on. It is a summary of socio-economic characteristics. *Highest education* is a categorical variable on the highest level of education achieved, ranging from "Year 11 and below" to "Postgraduate". Finally, *Parents divorced* is a dummy variable equal to 1 if the parents are divorced, and 0 otherwise.

Table 1. Descriptive statistics

	Mean
Duration[1,2]	
Of unions	4.67
	(4.52)
Of marriages	6.77
	(4.47)
Of pure cohabitations[3]	3.87
	(3.37)
Of premarital cohabitations	3.00
	(2.10)
Unions[2]	
Union involves marriage	0.40
Marriage started with cohabitation	0.26
Union is purely cohabitational	0.60
Partners' Characteristics[2]	
Birth cohort	1980†
Remoteness of area	0=*major city*†
Relative disadvantage	4=4^{th} *decile*†
Highest education	5=*Certificate III or IV*†
Parents divorced	0.17
No. distinct unions	3,963
No. individuals	5,350

Note. The table includes only observations on couples whose relationship began outside of NSW and after the year 2000. The table reports mean and standard deviation (in parentheses) for the variables used in the analysis. The data is right-censored in year 2017. This implies that a union that is observed only cohabiting in the sample but will get married in 2021 is counted as purely cohabitational. Widowed single individuals are counted as married.

[1] Includes durations equal to zero, i.e. unions lasting less than a year.
[2] Unions are counted at the individual level. A couple counts as two unions.
[3] Cohabiting couples who never got married in the sample.
† Most frequent category (mode).

4 Selection Channel

4.1 Identification

The source of identification is the Family Law Amendment Act 2008, since it is exogenously passed in a particular moment in time. In an ideal setting, I would identify the selection effect by comparing unions formed just before the reform with unions formed just after. The unions in the two groups *live* under the same legal regime, that being the new one. They are only different in the legal regime under which their union *began*: the control group started under the old regime, the treated group under the new one. In the analysis, I keep all the unions that began in a 3-year window from either side of the reform's year 2009, i.e. between 2006–2011. The unions that began between 2006-2008 are assigned to the control group and the ones that began between 2009-2011 are assigned to the treatment group. The choice of a 3-year window is made in order to increase the sample size[2].

4.2 Sample Construction

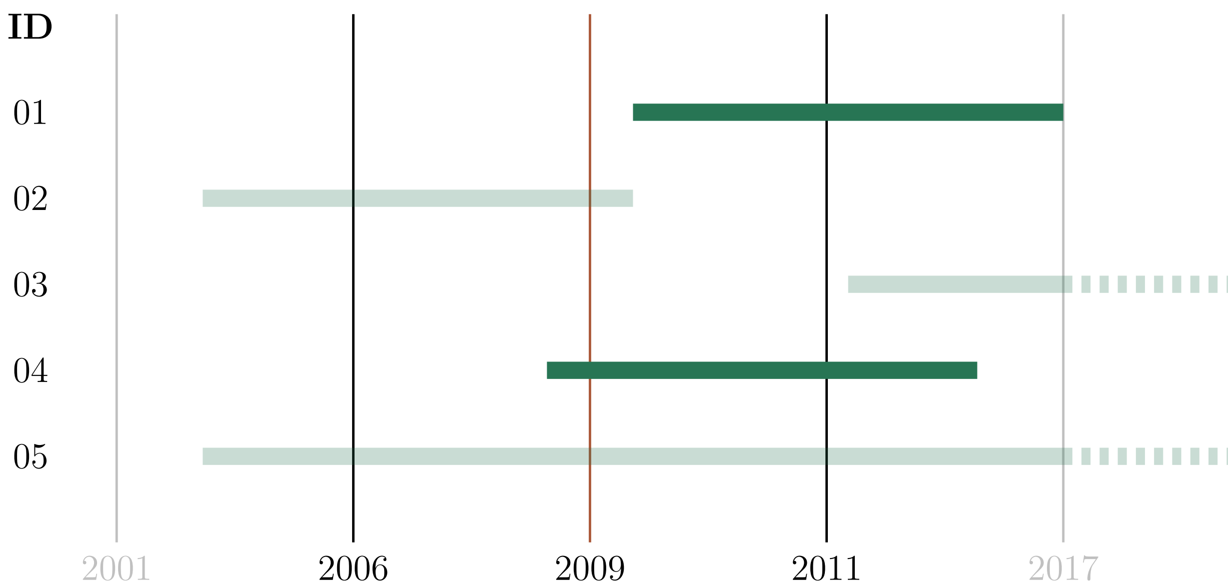


Figure 1. Visualisation of the selection channel sample

As visualised in Figure 1, the sample for separating the selection channel is constructed by only keeping unions that started in a 3-year window around the calendar year in which the reform became active (the darker lines). These unions are then followed over time. Those unions that started before 2009 are *always* in the control group, even after 2009, and those started during or after 2009 are always in the treatment groups, even after 2011.

[2]As shown is Section 4.6 results are robust to the use of a 2-years window

Furthermore, to better separate selection and incentive effect, the observations on the first two periods of all unions are dropped. Indeed, unions that began in 2007 remain in the old legal regime approximately two years before entering the new regime in 2009, during their third year. Unions started in 2006 remain for two years. In other words, those observations need to be dropped in order to exclusively compare unions *existing* under the same regal regime, which is the object of this section. Conversely, if those initial periods were kept, the comparison would include pre-reform unions during the old legal regime, thus introducing bias in the selection effect estimates.

In this sample, the unions that began as a marriage are dropped, as they are not affected by the reform on cohabitations. All the remaining unions start as a cohabitation, which then either becomes a marriage or does not. Each individual can only have one union at a time but more than one union over time.

4.3 Empirical Specification

To estimate the selection effect of an increase in the exit costs of cohabitation on union stability, I estimate the following regression equation using a linear probability model (LPM)[3]:

$$Pr[S_{j+1} = 1 | S_j = 0, D, \boldsymbol{X}] = \alpha_0(j) + \alpha_1(j)D + \boldsymbol{\beta X} \tag{1}$$

where S is a separation dummy, equal to 1 if the union ended at time j and 0 otherwise; D is a treatment dummy, equal to 1 if the union started strictly after 2008, 0 otherwise; it is flexibly interacted with period variable j, using a third degree polynomial, so that $\alpha_k(j) \equiv \gamma_{0k} + \gamma_{1k}j + \gamma_{2k}j^2 + \gamma_{3k}j^3, k \in \{0, 1\}$. Finally, $\boldsymbol{X}$ is a vector of birth cohort dummies (one per decade). Standard errors are clustered at the union level.

4.4 Results

Estimating Equation 1 on unions formed three years before and after the reform, I find that new unions are more stable (Figure 2). This is consistent with the hypothesis that the higher expected costs introduced by the reform deters the lowest quality matches from starting a cohabitation (see Section 2). Between the third and eighth year of their relationship, new unions are less likely to separate compared to the old unions formed in the three years before the reform. In particular, after the fifth year the effect is significant at 5% or less, with per-period point estimates around 2% (as can be calculated from Table 2). In other words, if we compare couples that lived both under the new legislation but formed under different ones, we find that the ones formed under laws imposing a higher cost of exit from cohabitation are more stable. This is consistent with the new law incentivising better matches on average.

[3]Results are virtually identical if a Logit model is used instead

Table 2. Selection effect: policy impact on new unions' per-period probability of separation

	(1) Separated
D	-0.0290
	(-0.14)
j	-0.129
	(-1.33)
$D \times j$	0.0180
	(0.14)
$j \times j$	0.0238
	(1.29)
$D \times j \times j$	-0.00432
	(-0.18)
$j \times j \times j$	-0.00143
	(-1.28)
$D \times j \times j \times j$	0.000273
	(0.19)
Birth cohort=1940	0.0445*
	(2.41)
Birth cohort=1950	0.0345***
	(3.73)
Birth cohort=1960	0.0372***
	(4.88)
Birth cohort=1970	0.0395***
	(5.88)
Birth cohort=1980	0.0402***
	(7.34)
Birth cohort=1990	0.0449***
	(4.50)
Constant	0.225
	(1.41)
Observations	7202

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table presents the LPM estimates of the impact of the Family Law Amendment Act 2008 on the probability of separating for new unions. The interaction of the treatment D with a third degree polynomial of the union's duration j allows for a compact but flexible specification of the policy impact.

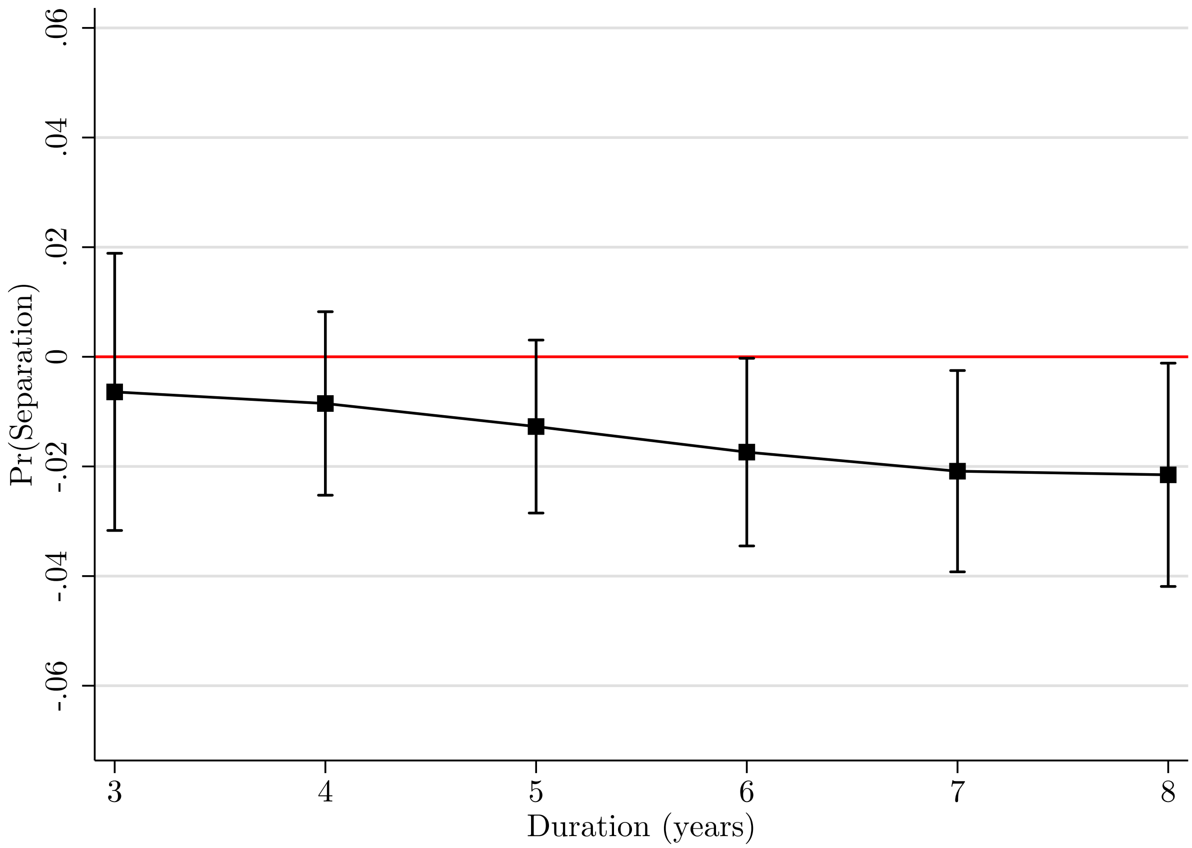


Figure 2. Selection effect: policy impact on new couples' per-period probability of separation

4.5 Separating the effect by union-type

This estimated effect on unions does not, by construction, separate between the impact of the reform on cohabitations-only unions and the impact on marriages that began as cohabitations. To do this, I estimate Equation 1, restricting the sample respectively (i) to cohabitation-only unions and (ii) to marriages that began as a cohabitation inside the 3-year window. When I look at unions which were pure cohabitations (Figure 3) I find a similar pattern as in the baseline analysis (Figure 2); after their fifth year, they are less likely to separate. The estimated effect is less precise but stronger, with new cohabitations 4% less likely to separate in period 6 compared to cohabitations that began before 2009, 5% less likely to separate in period 7 and 8% less likely to separate in period 8. Taken in isolation, these estimates are limited in that they only use information from cohabitations that either ended or that are censored, while discarding information from the cohabitations that became marriages.

If instead I restrict the baseline sample to marriages that began as cohabitations within the 3-year window, I still find a negative and significant 2% reduction in the probability of separation of new unions, but in their second and in third year of marriage (Figure 4). These estimates are again more limited than the baseline ones, particularly because I compare marriages that have premarital cohabitations of diverse lengths. This heterogeneity does not allow me to isolate the pure selection effect by dropping some initial periods, as done in Section 4.3. However, the sign and magnitude of these estimates suggests that the increase in match quality (as measured by stability) gained at the cohabitation stage has a positive impact on union stability also in the first years of marriage.

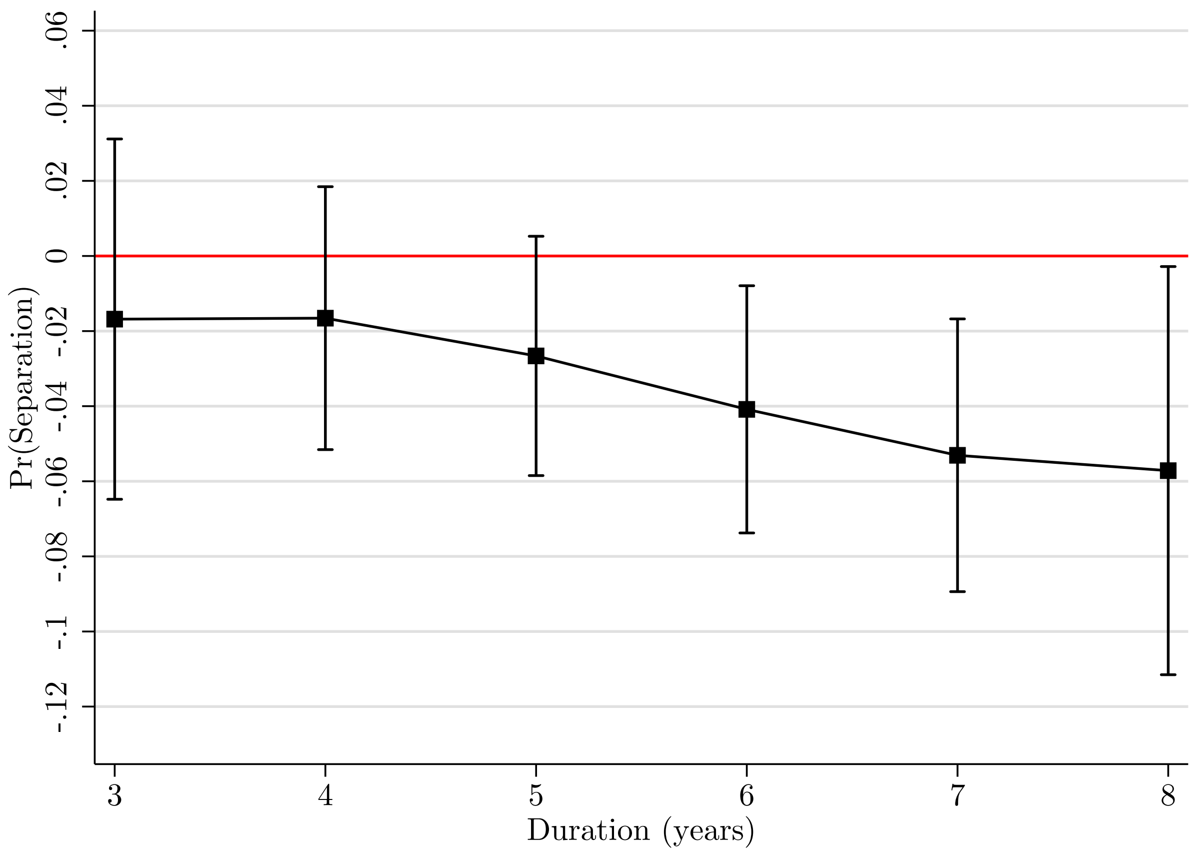


Figure 3. Selection effect: policy impact on purely cohabiting unions

4.6 Robustness Checks

Our policy of interest changes the legislation on cohabiting couples only, hence it should not affect unions which started as a marriage, without premarital cohabitation. Indeed, I find that the reform has no statistically significant selection effect on individuals who got married without cohabiting first (Figure 5). This is further evidence that the baseline results (Section 4.4) capture the causal effect of the policy of interest, as opposed to capturing some other general shock to relationship stability.

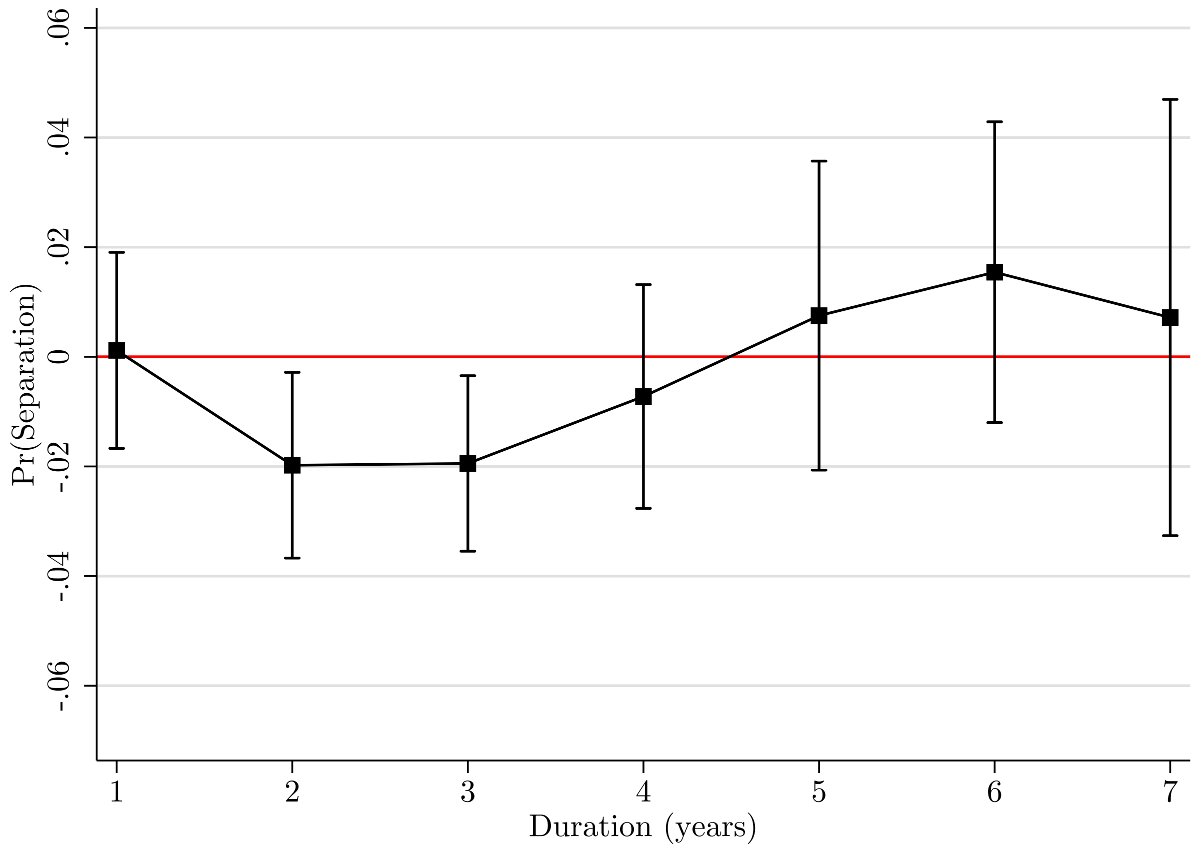


Figure 4. Selection effect: policy impact on marriages started as cohabitations

Adding covariates that are correlated with divorce can be useful for balancing the sample[4] (see the list in Section 3.2). I add them to Equation 1 and find that the estimated selection effect has a similar (per-period) magnitude compared to the baseline, but is now significant at 5% even in periods 4 and 5 (Figure 6). This suggests that the reform would produce more stable matches even if control and treatment group were identical in their observable characteristics predicting divorce.

Lastly, I test whether my findings can be replicated using placebo reform years (2003, 2004, 2005, 2006), in years antecedent to the 2008 reform. As expected, I find no selection effects in the previous years (7). This evidence suggests that the baseline estimates do capture the selection effect of the 2008 Family Law Amendment Act.

[4]Here I am using the full selection channel sample, which includes all the types of unions.

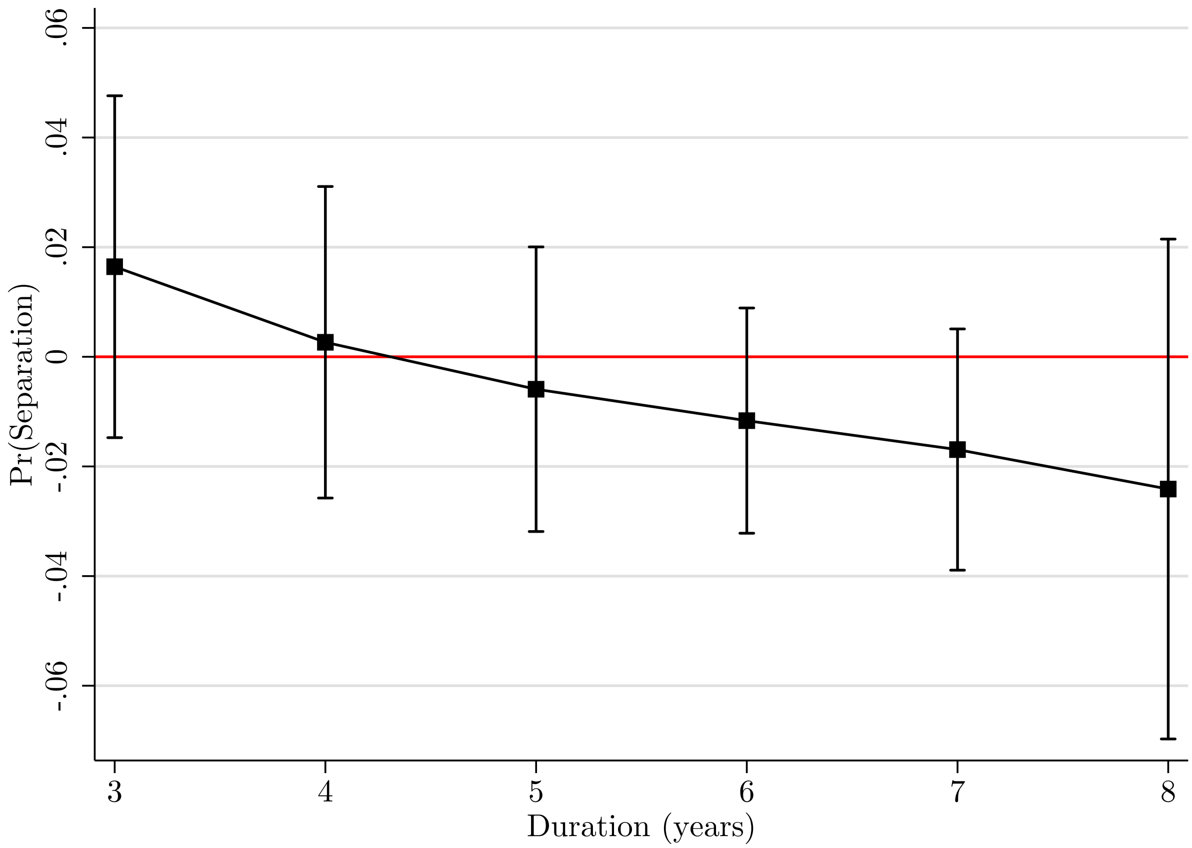


Figure 5. Selection effect: policy impact on purely marital unions

5 Incentive Channel

5.1 Identification

Imagine an experiment on cohabiting couples. All couples start cohabiting under a low-exit-cost regime — the old legal regime. After j years, a random group is assigned to a high-exit-cost regime: this is the treatment group. The remaining cohabitations form the control group and remain under the low-exit-cost regime. Estimand is then the effect of the reform on couples that are in their j^{th} period in 2009, when the reform becomes active. In other words, the aim is to compare unions that are affected by the reform since their j^{th} period with couples which were not affected in their j^{th} period. Both treated and control unions are already formed when the law changes, so they experience a change in their incentive structure *during* their relationship. Hence changes in their behaviour are attributed to changes in their incentive structure, in particular in their exit cost from cohabitation.

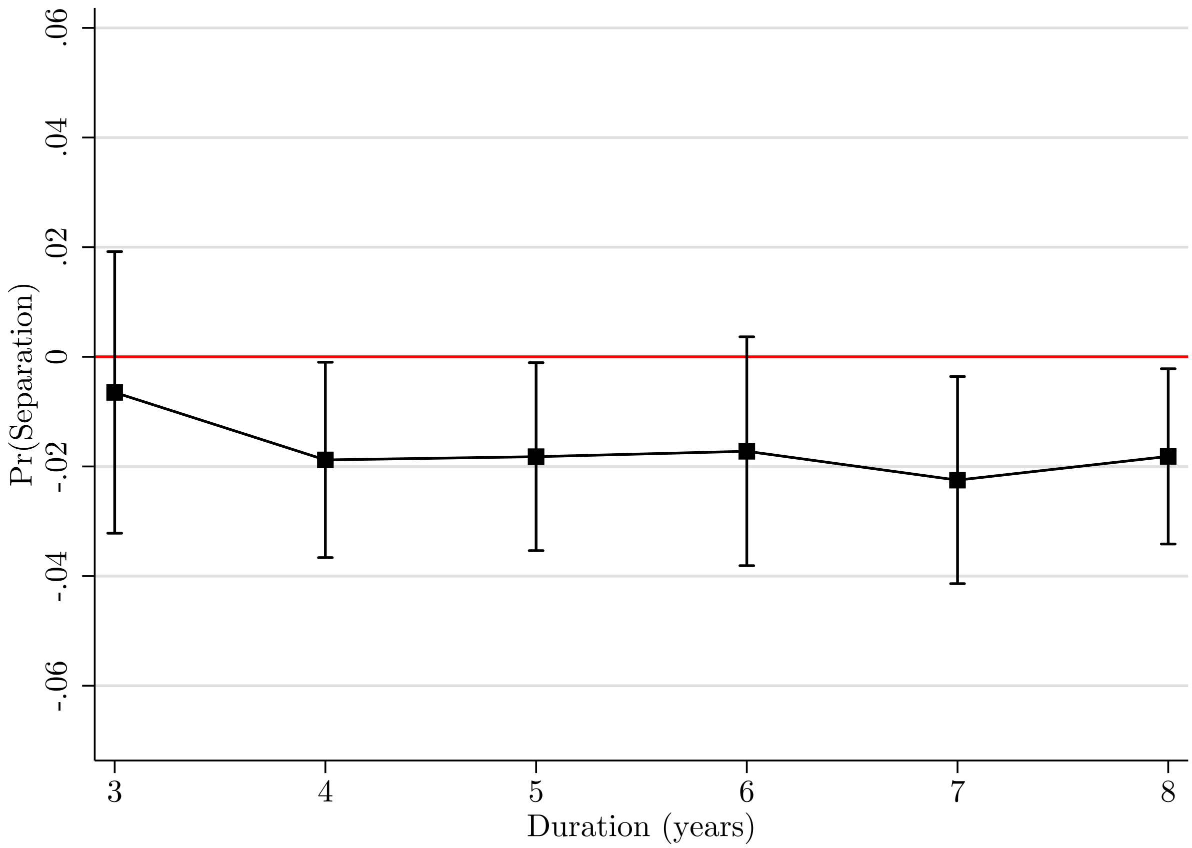
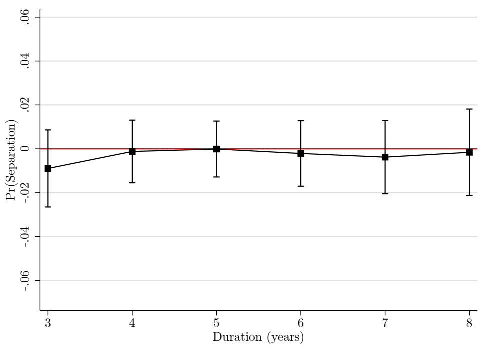


Figure 6. Selection effect: policy impact controlling for divorce predictors

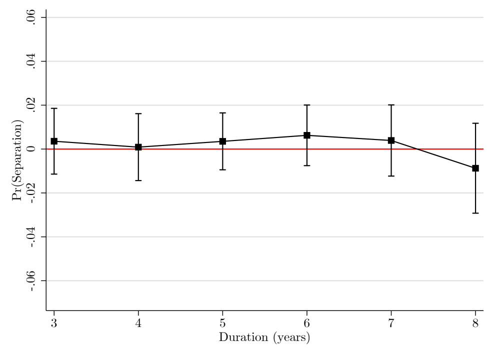
5.2 Sample Construction

My analysis in this section approximates such experiment by keeping only those unions that were in their j^{th} period within a 3-year window from the reform year, 2009. This is called a rolling window approach, because the 3-year window rolls back to "older" relationships[5] the higher the value of j. This can be visualised in Table 3, where, as the union's duration j increases, the coloured calendar years inside the window remain constant, while the years in which the union began decrease by one for each additional period of duration j. Notice that observations of cohabitations in period 1 and 2 are not used. Take period 2 as an example: it would be impossible create a subset of cohabitations that were in their second year in 2009 and that started in 3 different years (as necessary in order to construct a 3-year window). Similar issues apply to observations such that $j > 5$.

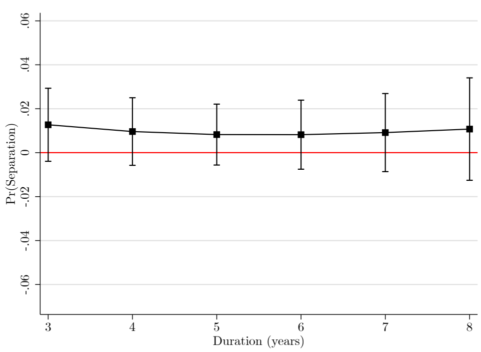
[5] Relationships that started before



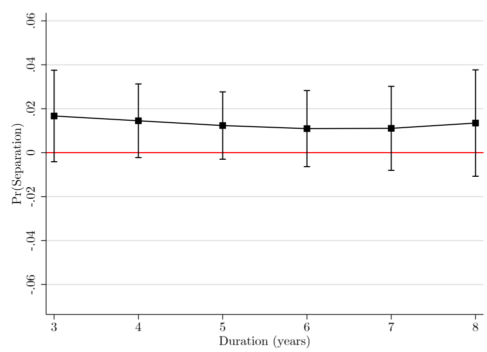
(a) 2003 placebo reform



(b) 2004 placebo reform



(c) 2005 placebo reform



(d) 2006 placebo reform

Figure 7. Selection channel: placebo effect

Table 3. Incentive channel: sample visualisation

year of start \ duration	3	4	5
2001	2004	2005	2006
2002	2005	2006	2007
2003	2006	2007	2008
2004	2007	2008	2009
2005	2008	2009	2010
2006	2009	2010	2011
2007	2010	2011	2012
2008	2011	2012	2013

Note. The table visualises how the sample was constructed. The values inside the table represent the current years. The 3-year window is coloured in grey for the control group and in brown for the treatment group. Each cell represent all cohabitations started in year of start s that in calendar year y were their j^{th} period of duration.

5.3 Empirical Specification

To estimate the effect of introducing the reform while an individual is in its j^{th} year of cohabitation, I estimate the following equation using the linear probability model (LPM)[6]:

$$Pr[S_{j+1} = 1 | S_j = 0, \bar{D}_j, M_j = M_{j+1} = 2, \boldsymbol{X}] = \beta_0(j) + \beta_1(j)\bar{D}_j + \boldsymbol{\beta_2 X} \tag{2}$$

where S is a separation dummy, equal to 1 if the cohabitation ended at time j and 0 otherwise; $\bar{D}_j$ is a treatment dummy, equal to 1 if the j^{th} period cohabitation takes place strictly after 2008, 0 otherwise; it is flexibly interacted with period variable j, using a third degree polynomial, so that $\beta_k(j) \equiv \gamma'_{0k} + \gamma'_{1k}j + \gamma'_{2k}j^2 + \gamma'_{3k}j^3, k \in \{0, 1\}$; M_j is a marital status categorical variable equal to 2 if a union is a cohabitation in period j. Finally, $\boldsymbol{X}$ is a vector of birth cohort dummies (one per decade). Notice also that unions cannot go from marital to cohabiting. Standard errors are clustered at the union level.

5.4 Results

Individuals who are in their third year of cohabitation in 2009 when the reform becomes active are 5% more likely to separate. Because the 2009 Family Law Amendment Act does

[6]Results are virtually identical if a Logit model is used instead

not define a cohabitation length after which it is classified a de facto relationship, it is impossible to predict a discontinuity around a specific threshold. However, it is also unlikely that a cohabitation that has lasted for several years would not be considered as a de facto. Given this lack of information, individuals might have relied on some rule of thumb, believing on average that any cohabitation longer than three years would not have been be considered a de facto relationship. If that were the case, an increase in separation rates in period 3 might come from partners in low quality matches who want to separate before they are treated as married. However, while New Zealand's Property (Relationships) Act 1976 sets the threshold for not being considered a de facto at the end of the third year of cohabitation, anecdotal evidence from Australia points towards a two-year one (Bryce, 2019). However, this is not part of the law and it could have emerged more recently. Indeed, if the two-year rule were followed immediately in 2009, we would have expected higher separation rates in the second year, close to the threshold for being considered a de facto.

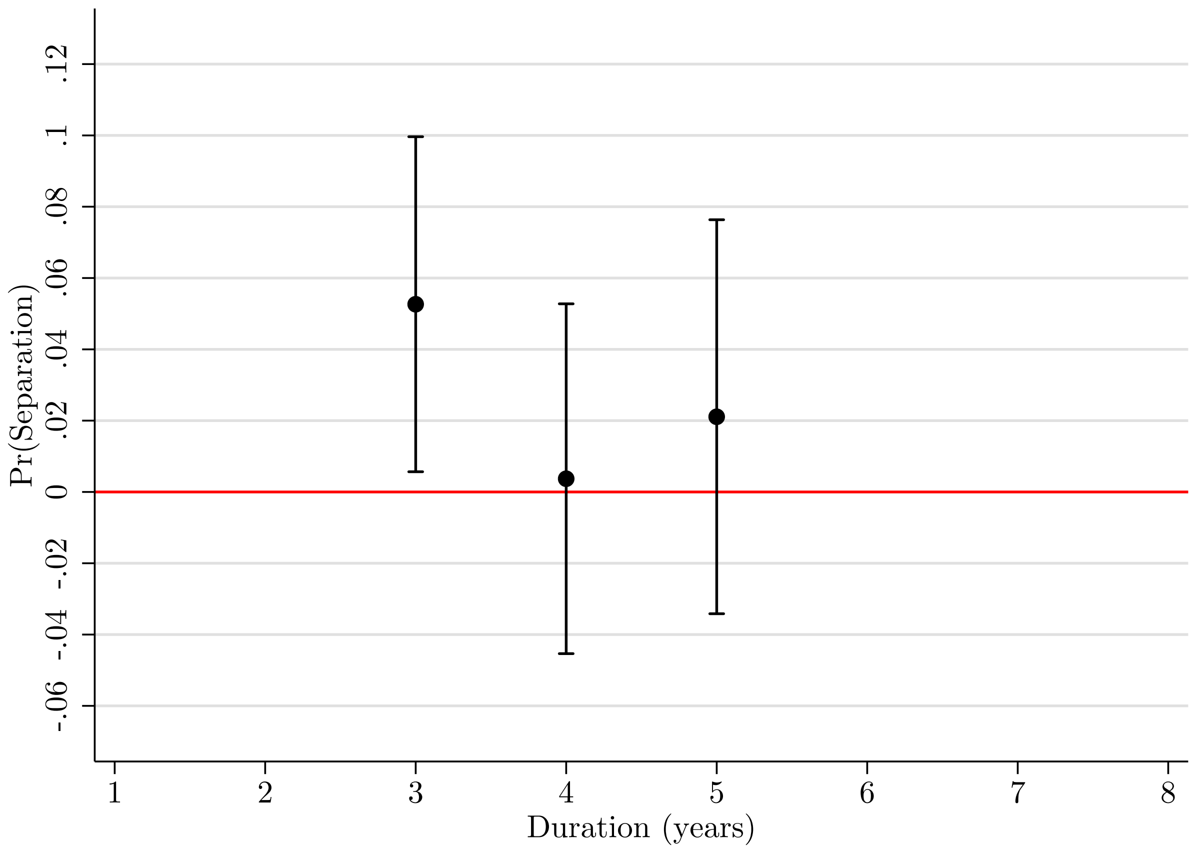


Figure 8. Incentive effect: policy impact on existing cohabitations' per-period probability of separation

5.5 Robustness Checks

To get closer to the ideal experiment, I restrict the window to 2 years, at the expense of the sample size. When the window is restricted to 2 years before and 2 years after the reform (2007-2010), the estimates hold, despite the loss in precision.

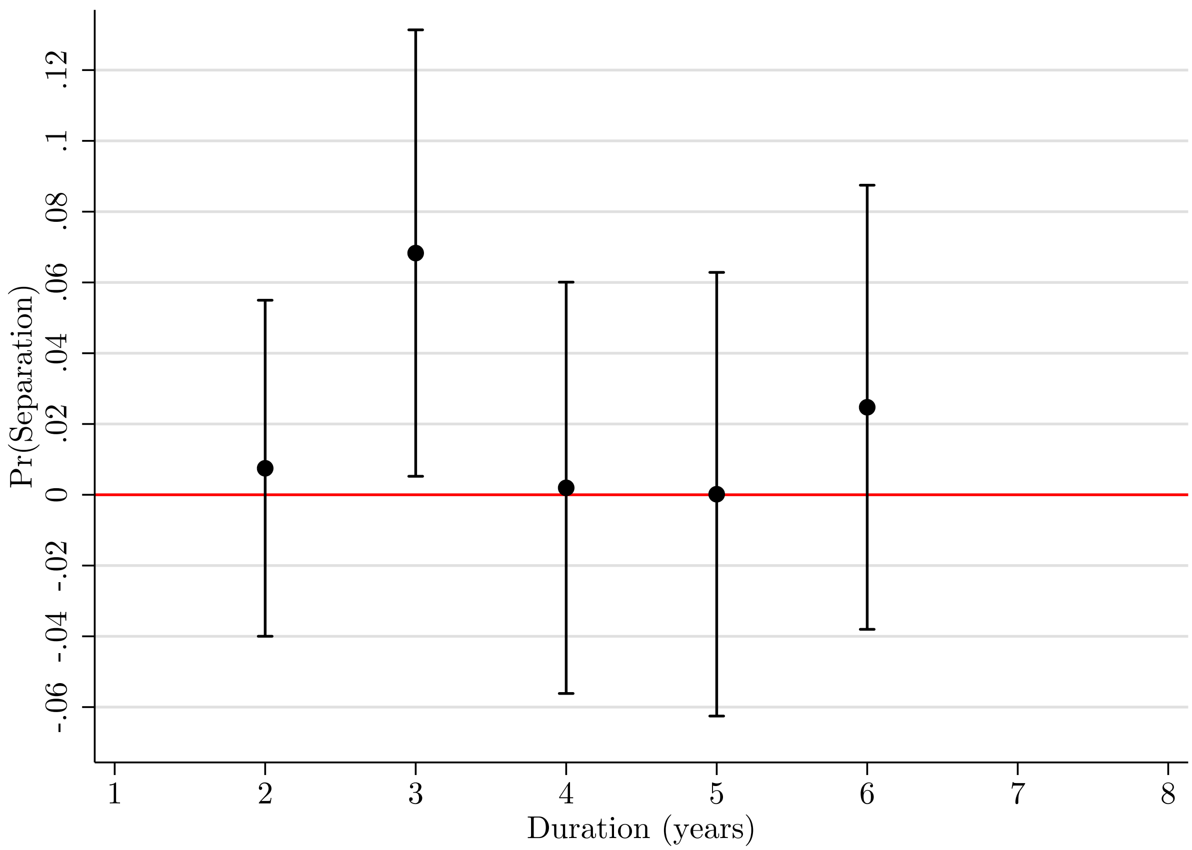


Figure 9. Incentive effect: policy impact on existing cohabitations using a 2-year window

Because the reform changes the incentives only for cohabiting couples, we should not observe any effect on married couples. To check if this is the case, I run a regression similar to Equation 2, but on a sample of married couples only:

$$Pr[S_{j+1} = 1 | S_j = 0, \bar{D}_j, M_j = M_{j+1} = 1, \boldsymbol{X}] = \beta_0'(j) + \beta_1'(j)\bar{D}_j + \boldsymbol{\beta_2'X} \tag{3}$$

where $M_j = 1$ for individuals married in period j of their union and the interpretation of the other parameters is unchanged from Equation 2. In line with the expected mechanism, I find that separation rates do not change for married couples when the reform becomes active in 2009 (Figure 10).

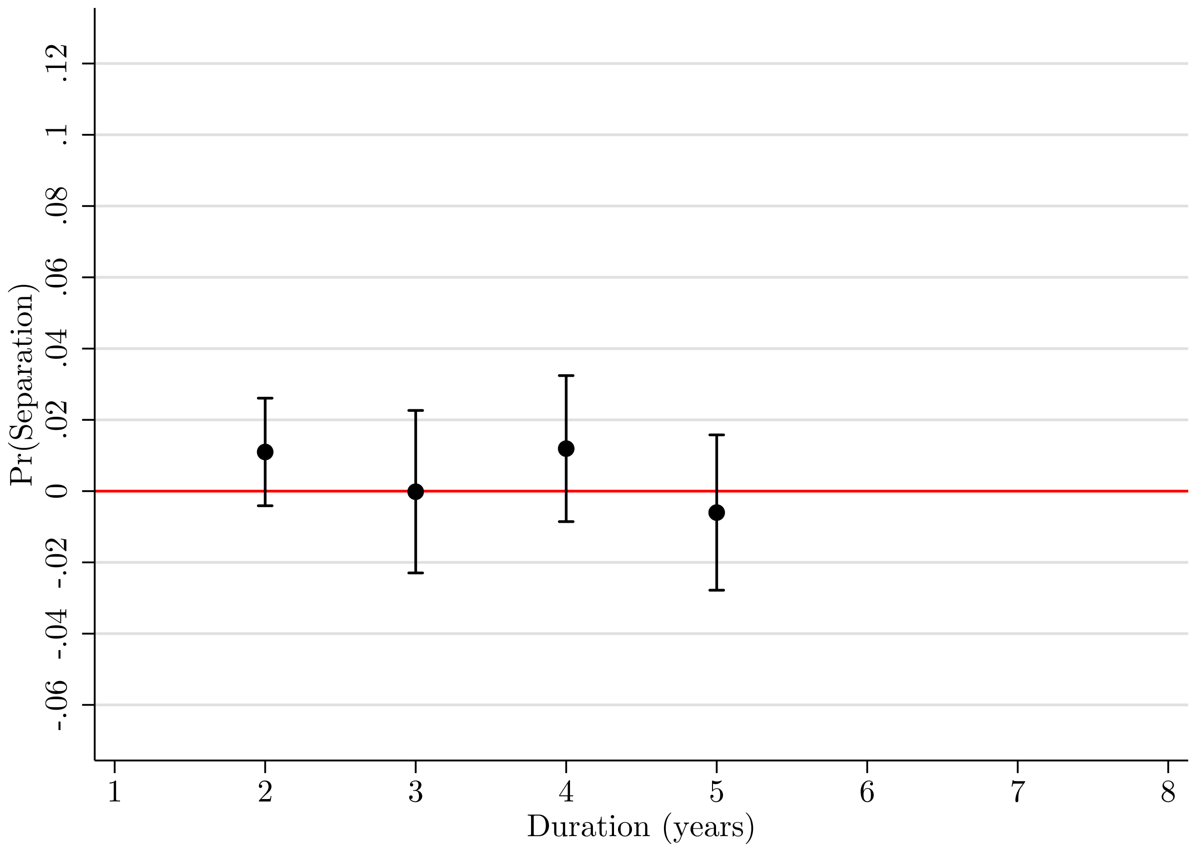


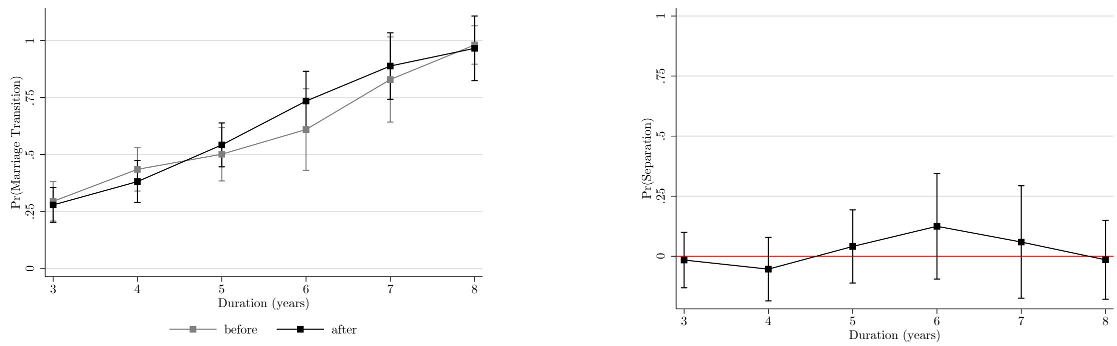
Figure 10. Incentive effect: policy impact on existing marriages' per-period probability of separation

6 Cohabitation-to-Marriage Transition

In this section, I test whether the reform has also shortened the phase of cohabitation for those partners who prefer to not get married immediately. Indeed, one might think that a reform that not only increases the cost of cohabitation, but makes it equal to marriage, could make shorter cohabitation phases preferable. Once the benefit of the flexibility provided by cohabitation is lost to the de facto status, it would be preferable to get married and enjoy the social benefits of it (Brien et al., 2006).

6.1 Sample Construction

To get some meaningful estimates on transition probabilities, I construct both a control and treatment group that include cohabitations that transitioned to marriage within the same number of periods. In particular, all cohabitations transition within their 8^{th} year of duration



(a) Transitioning before and after

(b) Policy effect on transitioning

Figure 11. Probability of transitioning from cohabitation to marriage

and start between 2006-2011, with a 6-year time window, as in Section 4.3. This implies that by period 8 all cohabitation will have transitioned to marriage both in the treatment and in the control groups (by construction). Purely marital unions are dropped from the sample.

6.2 Empirical Specification

I use a specification identical to Equation 1 to test whether premarital cohabitation has shortened after the reform:

$$Pr[M_{j+1} = 3 | M_j = 2, S_j = 0, Z = 1, \boldsymbol{X}] = \alpha_0''(j) + \alpha_1''(j)D + \boldsymbol{\beta''X} \tag{4}$$

where $M = 2$ if union is cohabitation and $M = 3$ if marriage; $Z = 1$ if a cohabitation eventually becomes marriage, while it is equal to 0 otherwise. The rest is identically defined as in Section 4.3.

6.3 Results

As Figure 11 shows, premarital cohabitations do not shorten after the reform. This means that the reform does not cause couples with a preference for a cohabitation phase to change the timing of their wedding. This result needs further study, as it can be easily shown that making terminating a cohabitation as expensive as divorce would induce a rational agent to prefer marriage to cohabitation in any model in which marriage provides an extra benefit relative to cohabitation. Restricting the window to 1 year, in order to estimate the effects for initial periods 1 and 2, provides further evidence of no effect (not shown).

7 New Cohabitations & Marriages

Given the claim that the observed increased stability found in Section 4 is due to the crowding out of low-quality matches[7], a decrease in new cohabitations is expected, or, similarly, a reduced probability for individuals (or dating couples) to start a cohabitation. A higher expected exit cost of cohabitation implies a higher match-quality threshold under which a non-cohabiting couple does not move-in. This implies that, ceteris paribus, finding a match good enough to justify starting a cohabitation should become rarer than before the reform.

The sample is constructed by using observations between 2003-2014 and keeping only the individuals aged between 16 and 60 in each year.

7.1 Empirical Specification

To estimate the probability of a non-cohabiting non-married individual to start a cohabitation, I estimate the following equation via LPM:

$$Pr[M_{j+1} = 2 | M_j = 1, \tilde{D}_t, \boldsymbol{X}] = \tilde{\delta}_0 + \tilde{\delta}_1 \tilde{D}_t + \tilde{\boldsymbol{\beta}} \boldsymbol{X} \tag{5}$$

where $M = 2$ if an individual cohabits and equals 1 if they single; $\tilde{D}_t$ is a treatment variable equal to 1 if $t \geqslant 2009$ and 0 otherwise, while $\boldsymbol{X}$ is again a vector of birth cohort dummies (one per decade), to control for generational differences. Standard errors are clustered at the individual level.

It can be noticed that this regression model is different from what seen in the previous sections. This is because here I am no longer looking at the expected duration before an event (separation) occurs. Instead, I am studying whether the yearly probability of an individual entering a union changes significantly after the reform is passed.

7.2 Results

Following the reform, the probability of starting a cohabitation does not change significantly (Table 4)[8]. This finding is inconsistent with most matching models, where higher exit costs crowd out the lower quality matches (see Matouschek and Rasul, 2008), which leads to a drop in new matches. To make these models consistent with my findings, one would have to make additional assumptions, for example higher exit costs making the search for a good match more efficient. In other words, assuming that high quality matches are scarcer than low quality matches, if the quality of new cohabitations increases but it remains as likely to start one, it means that the search for a match has improved on some level.

[7] Which is caused by the higher exit costs of cohabitation

[8] This result is robust to the use of a two-way fixed effect model.

Table 4. Probability of starting a cohabitation

	(1) New cohabitation
D	-0.00260
	(-1.73)
Birth cohort=1950	0.0101***
	(3.53)
Birth cohort=1960	0.0194***
	(6.68)
Birth cohort=1970	0.0372***
	(12.50)
Birth cohort=1980	0.0509***
	(17.91)
Birth cohort=1990	0.0340***
	(10.48)
Constant	0.0125***
	(5.16)
Observations	82145

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. This table presents the LPM estimates of the impact of the Family Law Amendment Act 2008 on the probability of starting a new cohabitation.

8 Conclusion

As the way in which people partner changes, governments are called to respond to these changes. While politicians often introduce regulations aiming at simply ratifying them, we know that this action is hardly neutral to the outcomes. Indeed, I show that increasing the expected costs of terminating a cohabitation leads to more stable unions. The positive effects are not exclusive to cohabitation stability, but spill over into the following marital phase.

This is inconsistent with standard models of cohabitation and marriage, which assume that the agent can always rationally rank the different marital states (single, cohabiting, married) – even during a relationship. Hence standard models do not allow the possibility that cohabitation might create distortions to rationality by strengthening the romantic attachment of the partners, even in situations in which they do not form a good match. The psychological literature as put forward such "cohabitation inertia" hypothesis, as detailed in Stanley et al. (2006). Stanley et al. (2006) claims that cohabitation makes one more likely to marry her partner compared to a no-cohabitation scenario. My results are consistent with cohabitation inertia, providing evidence that policies improving the stability of cohabitation improve the stability of marriage too. This particularly applies to countries such as Australia where marriage is mostly preceded by a period of cohabitation.

Other findings are also difficult to reconcile without departing from neoclassical assumptions. First, the duration of premarital cohabitation is not affected by the reform. On the contrary, once marriage and cohabitation are equalled and cohabitation loses its flexibility, premarital cohabitation time for new couples is expected to drop. Secondly, the number of new cohabitations remains stable after the reform, while all standard models predicts that it should change with the change in exit costs (Matouschek and Rasul, 2008).

Finally, I find that existing cohabiting unions affected by the reform since their third year are 10% more likely to separate. This is consistent with the idea that the reform has an effect on match quality. The lowest quality existing cohabitors do not find worthwhile to maintain their union under the new exit cost regime, so they break up attempting to escape it.

From a policy perspective, my findings imply that governments should carefully consider how they decide to regulate cohabitation. In particular, making cohabitation a choice with important legal and economic ramifications can help to promote more stable households, to the benefit of all members.

From a research perspective, this paper shows how cohabitation laws can provide a precious opportunity to study the causal dynamics at the heart of household formation and dissolution. Establishing these causal links is the first step on the one hand to identify which models best explain these dynamics and on the other hand to update such models as to enable them to reproduce patterns observed in the data.

References

ABS (2011). Socio-economic indexes for areas (seifa). [Cited on page 5.]

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy*, 81(4):813–846. [Cited on page 1.]

Becker, G. S. (1974). A theory of marriage: Part ii. *Journal of Political Economy*, 82(2, Part 2):S11–S26. (Not cited.)

Brien, M. J., Lillard, L. A., and Stern, S. (2006). Cohabitation, marriage, and divorce in a model of match quality*. *International Economic Review*, 47(2):451–494. [Cited on page 19.]

Bryce, B. (2019). In a de facto relationship? it won't save you from the cost of a divorce. [www.abc.net.au; Updated 6-June-2019]. [Cited on page 17.]

Chiappori, P., Iyigun, M., Lafortune, J., and Weiss, Y. (2017). Changing the Rules Midway: The Impact of Granting Alimony Rights on Existing and Newly Formed Partnerships. *The Economic Journal*, 127(604):1874–1905. [Cited on page 3.]

Friedberg, L. (1998). Did unilateral divorce raise divorce rates? evidence from panel data. Working Paper 6398, National Bureau of Economic Research. [Cited on page 2.]

Hewitt, B., Baxter, J., and Western, M. (2005). Marriage breakdown in australia: The social correlates of separation and divorce. *Journal of Sociology*, 41(2):163–183. [Cited on pages 2 and 5.]

Jose, A., Daniel O'Leary, K., and Moyer, A. (2010). Does premarital cohabitation predict subsequent marital stability and marital quality? a meta-analysis. *Journal of Marriage and Family*, 72(1):105–116. (Not cited.)

Lee, J. Y. and Solon, G. (2011). The fragility of estimated effects of unilateral divorce laws on divorce rates. Working Paper 16773, National Bureau of Economic Research. [Cited on page 2.]

Matouschek, N. and Rasul, I. (2008). The economics of the marriage contract: Theories and evidence. *The Journal of Law and Economics*, 51(1):59–110. [Cited on pages 1, 2, 3, 21, and 23.]

Mill, J. S. (1994). On the definition and method of political economy. In Hausman, D., editor, *The philosophy of economics: An anthology*, chapter 1, pages 49–51. Cambridge University Press, Cambridge. [Cited on page 1.]

NSW Government (1984). De facto relationships act 1984 no 147. [Cited on page 4.]

Parliament of Australia (2008). Family law amendment (de facto financial matters and other measures) act 2008 no 115.
`https://www.legislation.gov.au/Details/C2008A00115`. [Cited on pages 2 and 4.]

Perelli-Harris, B., Berrington, A., Sánchez Gassen, N., Galezewska, P., and Holland, J. A. (2017). The rise in divorce and cohabitation: Is there a link? *Population and Development Review*, 43(2):303–329. [Cited on pages 1 and 2.]

Stanley, S. M., Rhoades, G. K., and Markman, H. J. (2006). Sliding versus deciding: Inertia and the premarital cohabitation effect*. *Family Relations*, 55(4):499–509. [Cited on page 23.]

Wolfers, J. (2006). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review*, 96(5):1802–1820. [Cited on page 2.]

Chapter II

Are municipalities in the red to go green? The Blue Flag case.

1 Introduction

As Buckley (2002) puts it "ecolabels and environmental accreditation are controversial topics in tourism". Indeed, the problem is more complex than it seems. On the one side, there is no consensus around a precise definition of "ecolabel", which has become an umbrella term associated with a variety of certification types: mandatory, compulsory, multidimensional, monodimensional, etc. (Jha et al., 1997). On the other side, the effectiveness of ecolabels cannot be generalised, because it depends on the features of each single one, together with how informed the consumers are on the specific issue the label tackles and how strong its reputation is (Thøgersen et al., 2010). Given these considerations, it should not come as a surprise that the evidence around the impact of ecolabels, and more particularly beach awards (McKenna et al., 2011), is mixed. In light of these external validity issues, it is important to focus on the assessment of popular ecolabels since they affect large areas and promise to bring together financial and wildlife conservation interests , to the benefit of all stakeholders.

In this paper, I focus on one of the most famous and widely-adopted ecolabel, the Blue Flag programme. In particular, I look at how being awarded for the first time[1] affects a municipality's balance sheet and its supply of collective tourist accommodation. I focus on first-time certifications and their temporary effects because I expect the maximum impact of the Blue Flag to be reached when the certification is new. Indeed, a first-time certification is celebrated locally by the municipality, nationally by traditional media and internationally on the Blue Flag programme's website (www.blueflag.global). Moreover, the effect is likely to be temporary in countries like Spain, Greece, France and Italy where a large fraction of coastal municipalities have already been awarded in the past decades. In particular, due to this Blue Flag crowdedness, the effect is likely attributable to the extra-ordinary positive media exposure a municipality receives after a first-time certification.

[1]Throughout the paper, 'municipalities awarded for the first time' is intended as 'municipalities that are observed passing from not-certified to certified for the first time within the sample'. Lack of data on Blue Flag certifications prior to 2002 prevents me from being sure that any in-sample first certification is the first certification.

I adopt a pooled event study approach using Italian data between 2002-2016 to evaluate the effects of a first-time Blue Flag certification on the awarded municipalities. I find that municipalities awarded with a Blue Flag for the first time significantly increase their revenues, while I find no evidence that the award leads to an increase in the supply of collective tourist accommodation supply. My findings also provide further evidence (see Creo and Fraboni, 2011; Pencarelli et al., 2016; Cerqua, 2017) that, as proposed by Zielinski and Botero (2015), the Blue Flag award is an opportunity for mayors to promote and enact environmentally-conscious infrastructural improvements.

The literature on environmental beach certifications in tourist destinations is either survey (McKenna et al., 2011), or focusing on tourism in developing countries (Blackman et al., 2014), or looking at tourists' flows (Blanco et al., 2009; Capacci et al., 2015; Cerqua, 2017), or a combination of these and it focuses on the Blue Flag programme. For an updated review on the Beach Certification Schemes and the Blue Flag literature, see Zielinski and Botero (2019).

The findings on the economic effects of the Blue Flag award are not unequivocally significant and positive. At a survey level, McKenna et al. (2011) finds that beachgoers in Ireland, Wales, Turkey and the USA do not choose to visit a beach based on beach awards. However, Blackman et al. (2014) reports that the Blue Flag certification led to "19 new hotels and 1628 new hotel rooms per year" in Costa Rica. In developed countries such as Italy (see Capacci et al., 2015; Cerqua, 2017), where most of the statistical analyses have been carried out, the effects seem much more moderate . As Zielinski and Botero (2019) have noted, some anecdotal evidence seems to have led researchers to think that achieving and maintaining a Blue Flag certification is costly (see Blackman et al., 2014), to the point that it might exceed the economic benefits (see McKenna et al., 2011). However, Pencarelli et al. (2016) report that most of the 2012 Italian recipients of the Blue Flag which had a dedicated budget for it[2] allocated only up to €5K annually.

Using a province-level (*circoscrizioni turistiche*) Italian dataset, Capacci et al. (2015), via a Generalised Method of Moments (GMM) estimation of a demand equation, find that only foreign tourists arrivals are significantly and positively affected by the Blue Flag, while domestic tourists are not affected. However, Cerqua (2017), using different province-level Italian data, finds no evidence of a positive impact of the Blue Flag award on the flow of international tourists and evidence of a positive and significant effect on the flow of domestic tourists, only when the certification comes with a wider sustainability policy. He takes a *reduced-form* approach and estimates the Blue Flag effect via synthetic control methods.

This lack of consensus, even within the same award and country, is in part due to the severe issue of self-selection affecting these environmental awards and eco-labels, for which application is voluntary (see Blackman et al., 2014). Crucially, the lack of information on those who "almost" won and those who "just" won an award – based on some objective criteria – prevents the researcher to compare the most similar units across winners and losers. Moreover, the Blue Flag is awarded through a point-based system, a design well-suited for policy evaluation studies. However, in the Blue Flag case, the NGO operating the

[2]54% of the Italian winners of the Blue Flag in 2012 had a dedicated budget for it.

programme has never released to researchers[3] neither this information nor information on which municipalities applied and did not win. This implies that the observed non-winners of the Blue Flag include both municipalities that never applied for the certification and those that applied but failed to obtain it. These shortcomings in the data, together with the fact that winners and non-winners are different from each other in important ways, make the construction of a counter-factual scenario, on which the impact evaluation estimation is based, rather difficult and assumption-heavy. Capacci et al. (2015) include in their dataset all the Italian coastal provinces, conditioning their tourism demand variables on a number of covariates that the literature has identified as determinants of tourism demand. They assume that by conditioning on those variables they are able to identify the tourism demand function and hence how tourism demand is affected by the award of a Blue Flag. Blackman et al. (2014) uses propensity score matching (PSM) to build a *control group* (of non-winners) which is then compared with the *treated group*. However, as King and Nielsen (2016) show, PSM often aggravate imbalance, inefficiency, model dependence, and bias, instead of improving on them. Finally, Cerqua (2017), starting from a dataset on tourism flows of all the 164 Italian coastal provinces, selects as the *treated group* 20 provinces that received an extra Blue Flag[4] between 2008-2012, while constructing the *control group* by creating synthetic non-awarded provinces. These are 150 convex combinations of non-awarded provinces with both similar characteristics and pre-award tourism history.

Because of the above-mentioned issues with constructing a valid control group, I construct my estimates via pooled event study analysis (see Cengiz et al., 2019), which relies only on information about the treated units. The idea behind this method is to define a time window around an event of interest (for example, from -2 years from the event to +2 years) and to use the observations outside of this window as a counter-factual for what is happening inside of it. Instead of assuming that the treated group and the control group are identical (in absence of the event/treatment), it assumes that the periods inside the window are identical to the periods outside of it (but for the event/treatment). This assumption is strengthened by the use of time and individual fixed effects, which control respectively for year-specific macroeconomic shocks and individual characteristics.

The other key choice the researcher has to make is what to define as 'treatment'. In the Blue Flag case, the choice is non-trivial, as the award, once won, might or might not be renewed each year, depending on whether the environmental standards are maintained. This could imply that consumers respond differently to a municipality being certified the first time versus one being re-certified for the n-th time, or even one re-obtaining a Blue Flag after losing it. However, this heterogeneity has not been recognised in the literature leading to treatment groups pooling some or all of these cases and hence potentially failing to capture the effect where it has occurred. I argue that the Blue Flag is mostly effective when first awarded, hence focusing on first-time winners. This also allows me to avoid this source of heterogeneity that pooling different types of Blue Flag implies.

My paper's contribution to the literature is threefold. First, by using pooled event study

[3]not even in anonymised version

[4]without having received an additional one in the previous year

analysis, it avoids the self-selection issue affecting the causal estimation of the effects of interest. Second, it uses data on a more disaggregated level compared to other developed-country studies. Third, it looks at balance sheet data, which is useful in in exploring the cost-effectiveness of the Blue Flag certification.

The rest of the paper is structured as follows. Section 2 gives a summary on what the Blue Flag certification is and what are its requirements. Section 3 reports what data is used and the sources from which they were taken. Section 4 explains the identification strategy and the regression specification employed. Section 5 presents the results, while Section 6 summarises the findings and derives some policy implications.

Figure 1: *Blue Flags in Italy, 2018. Map's source: carto.com.*

2 The Blue Flag Programme

The Blue Flag is a programme operated by the Foundation for Environmental Education (FEE), based in Copenhagen. It started as a European programme in 1987 and was then extended to extra-European countries in 2001, so that it is now operating in 49 countries. It is a label awarded to the beaches, marinas and eco-boats that meet a number of environmental, educational, safety- and accessibility-related criteria. In 2017, in Italy, beaches from 230 municipalities received the award. The criteria are verified each year and thus have to be maintained over time to retain the Blue Flag. The mission of Blue Flag is (FEE, 2017):(i) promote and participate in environmental education programmes for the users of beaches, marinas and eco-tourism boats; (ii) implement sound safety and environmental management systems; (iii) monitor environmental conditions to reduce the impact of human activity at the beaches, marinas and eco-tourism boats; (iv) commit to partnerships and collaborative action to promote the sustainable development of tourism. In 2017, 4423 Beaches, Marinas and Eco-tourism Boats in 49 countries (FEE, 2017) featured the Blue Flag, making one of the most successful eco-label worldwide.

3 Data

Several sources are used to construct the dataset. The tourism capacity data between 2002-2016 were taken from ISTAT's website (*Capacity of tourist accommodation establishments*). The municipal balance sheet data for the years 2004-2014[5] were kindly provided by Openbilanci.it, a DEPP and Openpolis project, cofunded by the European Union. Balance sheet information relative to the years after 2014 was not used due to radical changes in municipal accounting system. The Blue Flag data, was partly shared by Capacci et al. (2015), partly scraped from FEE's website[6], the NGO behind this certification. It ranges between 2000 and 2016 and the beach level data have been aggregated at the municipality-level. Table 1 presents the outcome variables used in the analysis.

All the dependent variables are transformed via inverse hyperbolic sine (IHS) transformation. The IHS transformation has the same properties and can be interpreted in the same way as a log transformation. This is because it approximates the logarithmic transformation but for very small values, as shown below:

$$IHS(Y) = log(Y + (Y^2 + 1)^{1/2}) \approx log(2Y) = log(2) + log(Y) \tag{1}$$

The advantage of the log transformation would that it is defined at $Y = 0$, while still allowing for interpreting the regression coefficients in Section 5 as semi-elasticities.

4 Methodology

4.1 Identification

The Blue Flag is a voluntary certification and this makes evaluating its impact challenging. As any voluntary policy or award, it is affected by a problem of *self-section*, in which the winners of the award are not picked at random within the population of interest, but are in fact a sub-sample of the population with specific characteristics. It is those specific characteristics that allow those municipalities (or units, more in general) to receive the award. A naive comparison between the mean outcomes of the two groups of winners (or *treated*) and non-winners (or *non-treated*) would therefore lead to a biased estimate. In principle, this issue could be ameliorated by comparing those municipalities which applied for the Blue Flag and did not get it with those which got it. Indeed, those applying and failing are more likely to be observationally similar to the winners compared to those which did not even apply.

[5]Balance sheets are always relative to the past year, so this dataset covers the municipal revenues and expenditures between 2003-2014

[6]www.bandierablu.org

Table 1: *Outcome variables descriptions*

Variable name	Description
Tourism services supply	
Total beds	Total beds available in collective accommodations
Hotel beds	Hotel beds available
Other beds	Extra-hotel beds available
Municipal Balance Sheet	
Total municipal revenues	Municipality's revenues from taxes and other sources
Coast-renting revenues	Revenue from renting state-owned properties
Physical capital	Investments in physical capital

This is true under the assumptions that both applicants and non-applicants are aware of the Blue Flag, of its requirements and find it desirable – reasonable assumptions given the long-standing popularity of this beach award. Furthermore, an even stronger identification of the causal impact of the Blue Flag could be provided by having access to the data on all the criteria based on which the municipalities are judged. This would allow the researcher to compare in a regression discontinuity design framework those municipalities that *almost* won the award with those that won it. Given that the FEE does not release any of this information neither to the public nor to researchers, other identification strategies must be adopted.

Figure 2 shows how different the treated and non-treated groups are. It plots a histogram of the probability of receiving the first Blue Flag for the never-certified and ever-certified[7] municipalities between 2002 and 2016, conditional on several census characteristics. Indeed, the two distributions are very dissimilar, the never-certified one being strongly skewed to the left while the ever-certified one being rather uniform. The issue is so severe that for a share of the treated municipalities there is no "comparable" non-treated one with whom comparing them. Because of this fundamental issue of constructing a valid control group, I construct my estimates using a pooled event study analysis (see Cengiz et al., 2019), which relies only on information about the treated units. The idea behind this method is to define a time window around an event of interest – for example, from -2 years from the event to +2 years after the event – and to use the observations outside of this window as a counterfactual for what is happening inside. Instead of assuming that the treated group and the control group are identical (in absence of the event/treatment), it assumes that the periods inside the window are identical to the periods outside of it (but for the event/treatment). This assumption is further strengthened by the use of time and individual fixed effects,

[7] i.e. those receiving at least one Blue Flag in the sample

which control respectively for year-specific macroeconomic shocks and constant individual heterogeneity. The pooled event study approach is a difference-in-differences approach. It consists in (i) comparing the observations inside the event window with the ones outside via period-specific dummy variables and (ii) taking the difference between the dummies' estimated parameters and the parameter associated with a chosen reference period, which must be inside the window and antecedent to the event. Because the pooled event study is a difference-in-differences, the parallel trend assumption must hold in order to identify the causal effect of the Blue Flag: absent of the Blue Flag certification, the outcome variable of interest in treated and untreated municipalities would move in parallel. A further required assumption is no heterogeneity in the effects across municipalities.

This approach avoids issue related to the use of propensity score matching, which have recently emerged in the literature (King and Nielsen, 2016). Blackman et al. (2014) uses propensity score matching (PSM) to build a *control group* of Blue Flag non-winners which is then compared with the *treated group*. However, as King and Nielsen (2016) show, propensity score matching often makes the comparison worse instead of better, by aggravating imbalance, inefficiency, model dependence, and bias. It also avoids the strong assumptions implied by a structural approach as in Capacci et al. (2015). (Capacci et al., 2015) include in their dataset all the Italian coastal provinces, conditioning their tourism demand variables on a number of covariates that the literature has identified as determinants of tourism demand. The causal interpretation of their estimates relies on the assumption that the tourism demand function is correctly specified.

4.2 Empirical Model

In order to study the effect of a Blue Flag certification over a range of municipal-level variables, I take a pooled event-study approach, following (Cengiz et al., 2019). I do this by choosing a a 5-year event window ranging between $[-3, 1]$ in annualised time, where $\tau = 0$ is the year when the Blue Flag is awarded, $\tau = -3$ is three years before the event and $\tau = 1$ is the year after. In the main specification, I look at how receiving a Blue Flag affects tourism services supply and the municipality's balance sheet, in particular during the award year and the following one. When the dependent variable is demand-side, I model the Blue Flag effect as temporary. In particular, I do this by estimating the following regression equation:

$$Y_{i,t} = \sum_{\tau=-3}^{1} \alpha_\tau I_{it}^\tau + \mu_i + \rho_t + u_{it} \tag{2}$$

where $Y_{i,t}$ is an outcome variable, while I_{it}^τ is an indicator variable which equals 1 if the Blue Flag was awarded τ years from calendar year t to municipality i, and 0 otherwise. I also control for both municipality, μ_i, and year, μ_i, fixed effects. After estimating Equation 2, I calculate the (percentage) change of the outcome variable between period -2 and period τ by normalising to α_{-2} the other α_τ coefficients, i.e. by subtracting $\alpha_\tau - \alpha_{-2}$. I choose $\tau = -2$ as the reference year in order to check for anticipation effects in period $\tau = -1$.

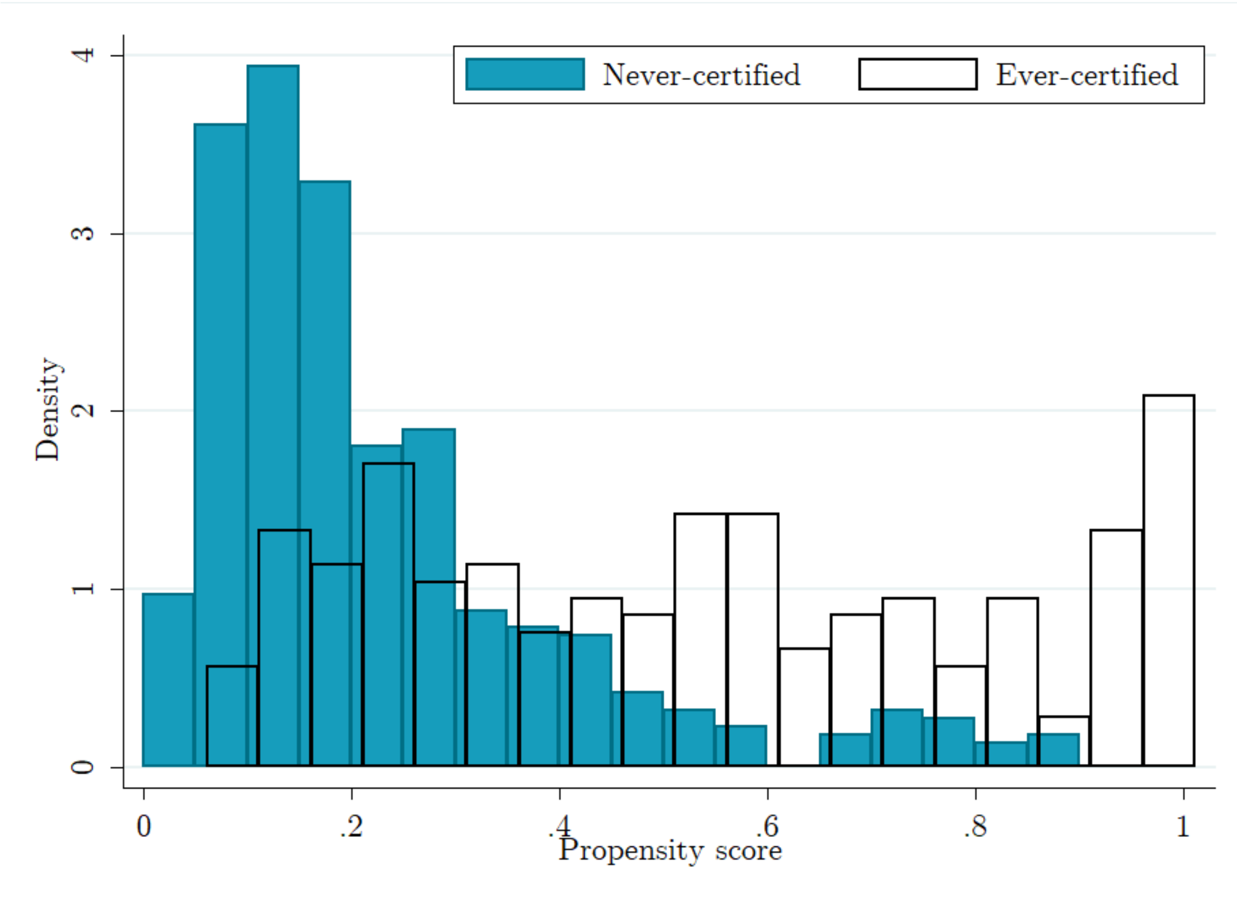


Figure 2: *Distribution of logit propensity scores for never-certified (blue bars) and ever-certified (empty bars) municipalities.*

4.3 Interpretation

The event of interest is the first Blue Flag certification a municipality wins for one of its beaches. It is expected to increase the economic activity in the associated municipality by boosting tourism. A significant increase of the dependent variable at $\tau \geqslant 0$ is evidence that the certification had an effect on it and non-zero effects at $\tau \leqslant 0$ might reveal pre-award patterns, such as anticipation effects and pre-existing positive trends. A casual interpretation of my estimates relies on the assumption that, conditional on municipality characteristics and common macroeconomic shocks, had the Blue Flag not been awarded, the outcome variable would have not significantly changed.

Table 2: ***Temporary effects of a Blue Flag certification.***

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Municipal Balance Sheet			*Supply of Beds in Collective Accommodations*			
Period	Revenues	Capital	Coastal Rent	Hotel	Extra-Hotel	Total	3+ Stars
-3	.085	.208	.910	-.015	.010	.001	-.033
	(.039)	(.130)	(.715)	(.036)	(.087)	(.044)	(.032)
	[.029]	[.111]	[.203]	[.679]	[.892]	[.988]	[.302]
-2	0	0	0	0	0	0	0
-1	.060	.209	1.863	.035	.047	.053	.109
	(.037)	(.124)	(.680)	(.033)	(.081)	(.041)	(.072)
	[.103]	[.092]	[.006]	[.292]	[.560]	[.197]	[.133]
0	.101	.354	1.596	.021	-.049	.036	.105
	(.037)	(.124)	(.681)	(.033)	(.081)	(.041)	(.095)
	[.006]	[.0046]	[.019]	[.525]	[.548]	[.376]	[.272]
1	.024	-.005	1.149	.022	.031	.041	.073
	(.039)	(.132)	(.725)	(.034)	(.083)	(.042)	(.102)
	[.539]	[.968]	[.113]	[.511]	[.705]	[.320]	[.474]

Note: The table reports the effect of a first Blue Flag certification on municipal balance sheets and tourist accommodation supply. Standard errors in parentheses, p-values in brackets.

5 Results

5.1 Temporary Effect

Figure 3 shows the temporary impact of being awarded a Blue Flag on municipal revenues. As reported in Table 2, column 1, winning the Blue Flag for the first time increases the revenues of the recipient municipality in the award year by 10% and the coefficient is statistically significant at 1%. A positive and statistically significant coefficient in period $\tau = -3$ could mean that the Blue Flag is particularly sought after by mayors as a way to promote the municipality following a decrease in revenues.

As Figure 4 shows, the supply of tourist accommodation does not respond to the Blue Flag certification. Indeed, the coefficients associated with $\tau = 0$ and $\tau = 1$ are not significantly different from zero when the dependent variable is, respectively, *number of hotel beds*, the *number of extra-hotel beds* (including AirBnB's, residences, campings, etc.) the *total number of tourist accommodation beds*, as columns 4 and 5 Table 2 show. The effect remains not statistically different from zero even when looking at the aggregate number of collective tourist accommodation's beds (Table 2, column 6. Given that the award does not

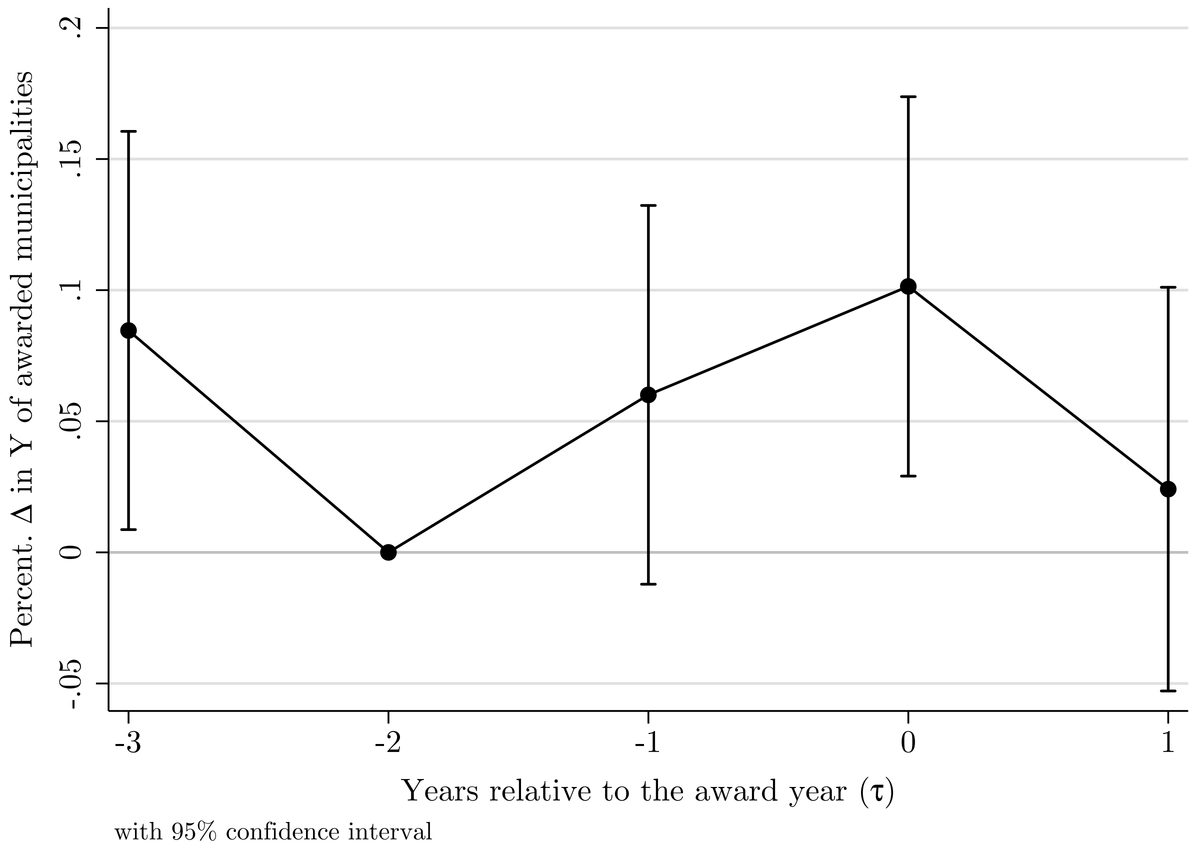


Figure 3: *Temporary impact of being awarded a Blue Flag on municipal revenues.*

significantly impact the overall *quantity* of the supply of accommodations, I study whether it has an affect on the *quality* of such accommodations. I do that by assessing the impact of the Blue Flag on the number of beds in hotels with 3 stars or more and I do not find evidence that accommodation supply is temporarily affected. However, it is not clear whether such supply variables should be considered rigid and hence the award impact on them modelled as permanent. This alternative specification is presented in Section 5.2.

Given the above results, the Blue Flag award seem to temporarily increase demand for tourism, as capture by municipal revenues, although not sufficiently to trigger an increase in supply too, which would be reflected in a positive change in the size of the accommodation industry. This result does not seem to hold in developing countries (see Blackman et al., 2014), where land available for construction is abundant and the tourism sector has a wider margin for growth.

In accordance with the literature (see Zielinski and Botero, 2019), my analysis of municipal balance sheet data shows that on average municipalities do not already meet the minimum requirements when applying for the Blue Flag, but rather invest in order to obtain it. Indeed, physical capital investments start increasing by +20% (significant at 10%) in the

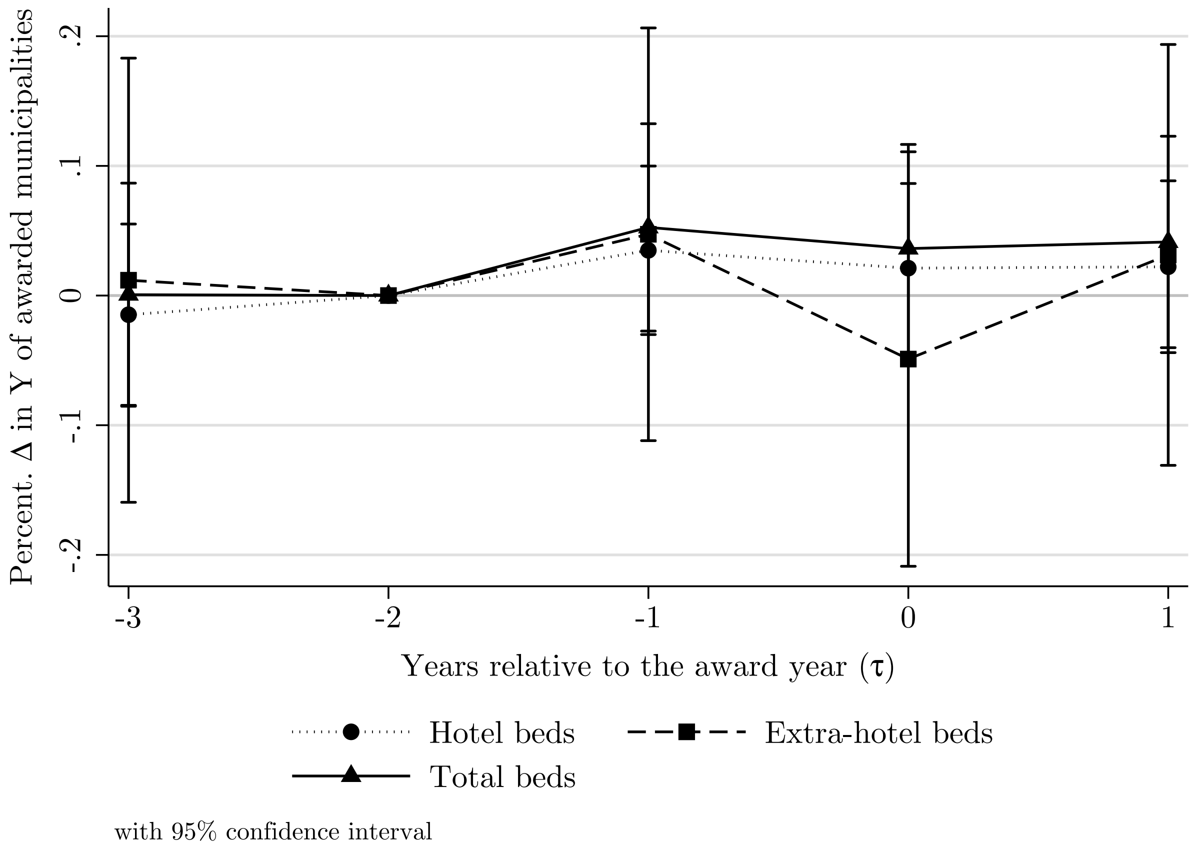


Figure 4: *Temporary impact of being awarded a Blue Flag on supply of collective accommodation.*

year before the award, peaking at +35% (significant at 1%) in the award year (see Figure 5). This is consistent with Zielinski and Botero (2019) which finds that the Blue Flag is (i) a "trigger of political will" that mayors use to effectively allocate resources and coordinate with the local businesses and expertise and (ii) the opportunity for pushing infrastructural improvements.

Revenues coming from renting municipal coastal or maritime areas (*concessioni demaniali*) for commercial use increase significantly and persistently by $10\% - 20\%$ since the year before the award (see Figure 6). Both this effect and the effect on municipal physical capital expenditure start in $\tau = -1$, suggesting the presence of anticipation effects. In other words, this evidence is consistent with winning municipalities investing substantially to meet the Blue Flag's infrastructural requirements while creating positive expectations around the outcome of the Blue Flag application. The private sector responds to this by increasing the commercial exploitation of the coastal areas, which in Italy are property of the government.

In summary, the Blue Flag's effect on municipal revenues is positive and sizeable. In Italy, the municipality budget must be balanced by law (Decreto Legislativo 18th August

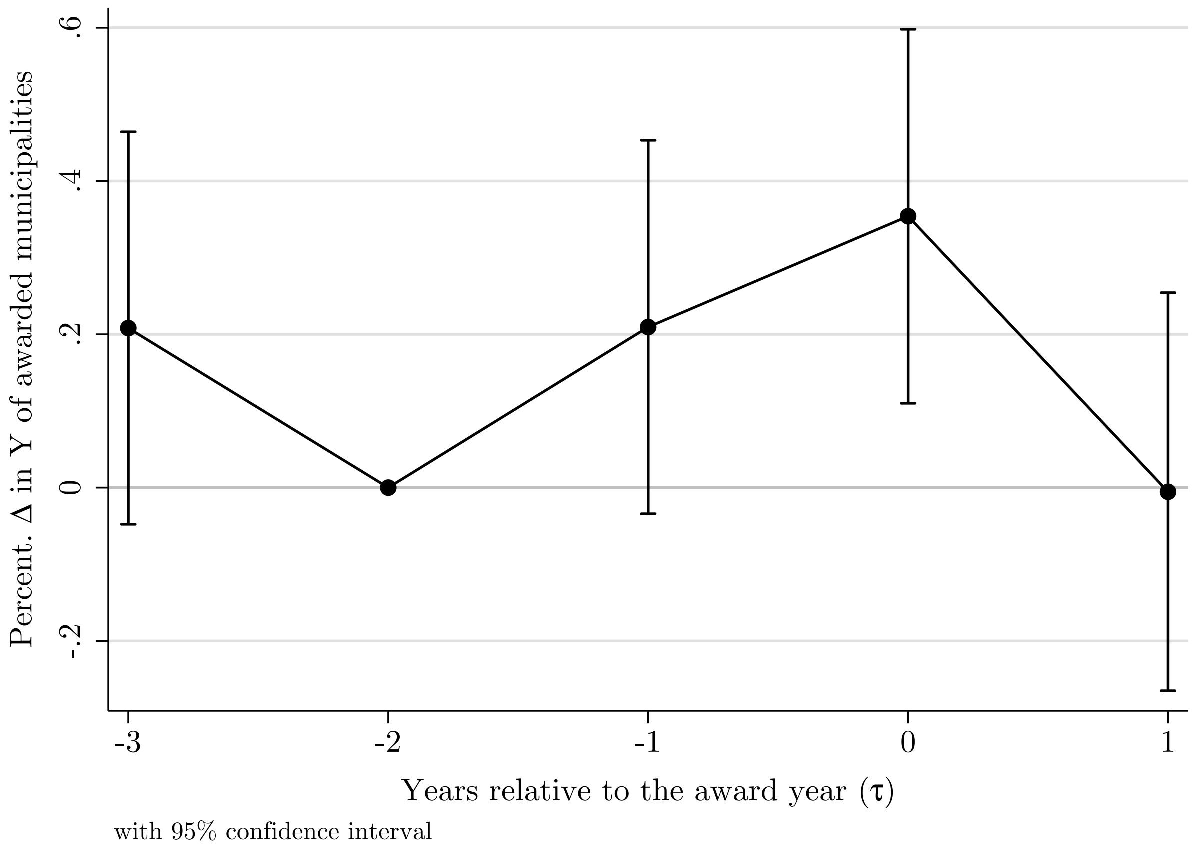


Figure 5: *Temporary impact of being awarded a Blue Flag on physical capital.*

2000, n. 267, art. 151), which implies that the Blue Flag's effect on municipal expenses and on municipal revenues are identical (and cancelling each other out). However, if we consider the increase in physical capital expenditure as a long-run investment that will permanently increase the population's welfare, then the net economic effect of a Blue Flag award is positive.

5.2 Permanent Effect

To test the robustness of my estimates, I estimate the same model (see Equation 2), but under the assumption that the effect of a Blue Flag award is permanent, rather than temporary. This is achieved by recoding the $\tau = 1$ dummy variable as equal to 1 when $\tau \geqslant 1$, i.e. in all periods strictly after the event year. It should be noticed that, if an estimated effect is similar when modelled as temporary to when it is modelled as permanent, this is evidence in favour of the temporary-effect specification. Indeed, modelling the effect as temporary means to calculate the counterfactual scenario using both pre- and post-event-window periods, while modelling the effect as permanent means using only the pre-event-window periods. If the

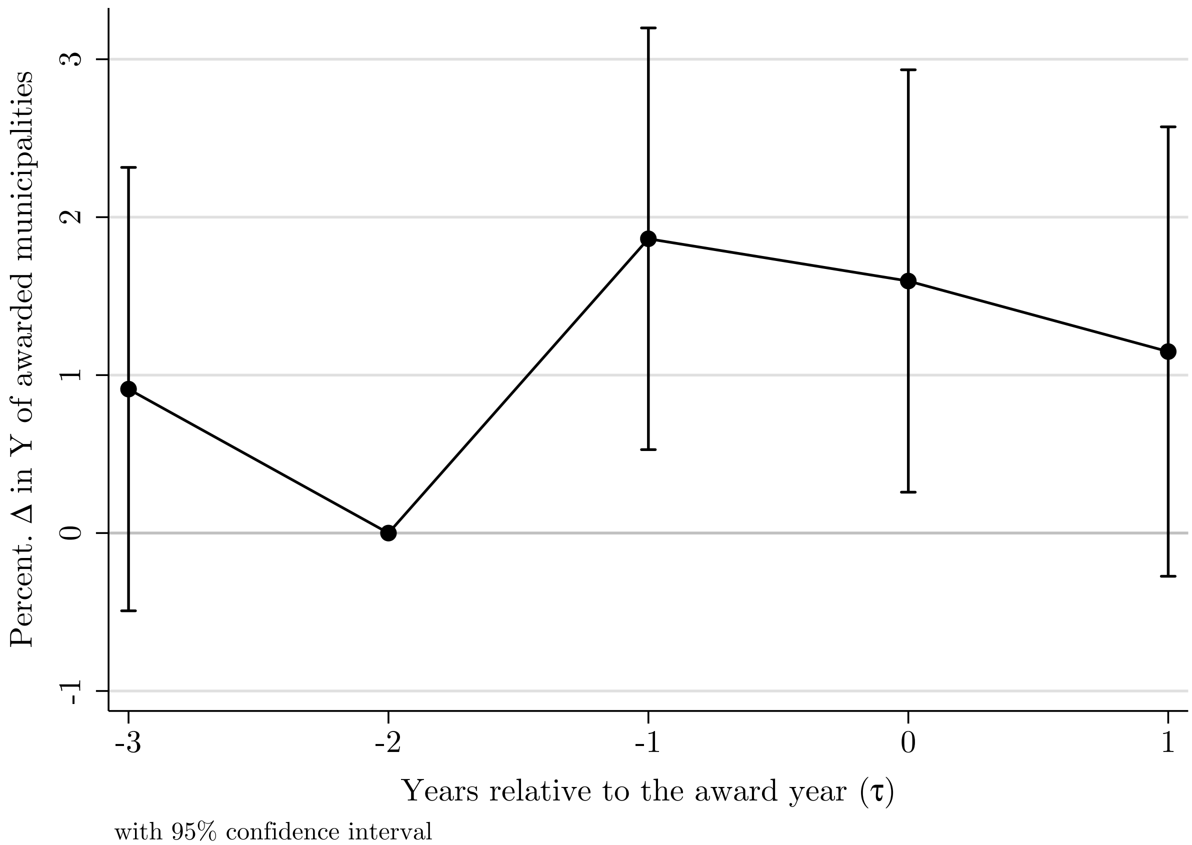


Figure 6: *Temporary impact of being awarded a Blue Flag on municipal revenues from renting coastal public property.*

effect is modelled as permanent while the true effect is temporary, the estimated effect will similar. However, if the effect is modelled as temporary while the true effect is permanent, the estimates will be biased, given that some periods used to calculate the counterfactual are affected by the award.

I find that modelling the Blue Flag effect as permanent rather than temporary leads to significant coefficients of the same sign and similar value relative to the ones estimated assuming a temporary effect – with one exception. Figure 7 shows as an example how the main results are robust to this assumption change.

The exception is found when analysing the Blue Flag effect on the supply of beds in hotels with 3 stars or more. As Figure 8 shows, the effect of receiving a Blue Flag for the first time becomes higher and significant at 10% in $\tau = 1$ when it is modelled as permanent. This, as explained above, is (weak) evidence of a permanent effect on this outcome variable, which is consistent with the view that the Blue Flag award increases demand for high quality accommodations and hotels respond by permanently increasing the its supply.

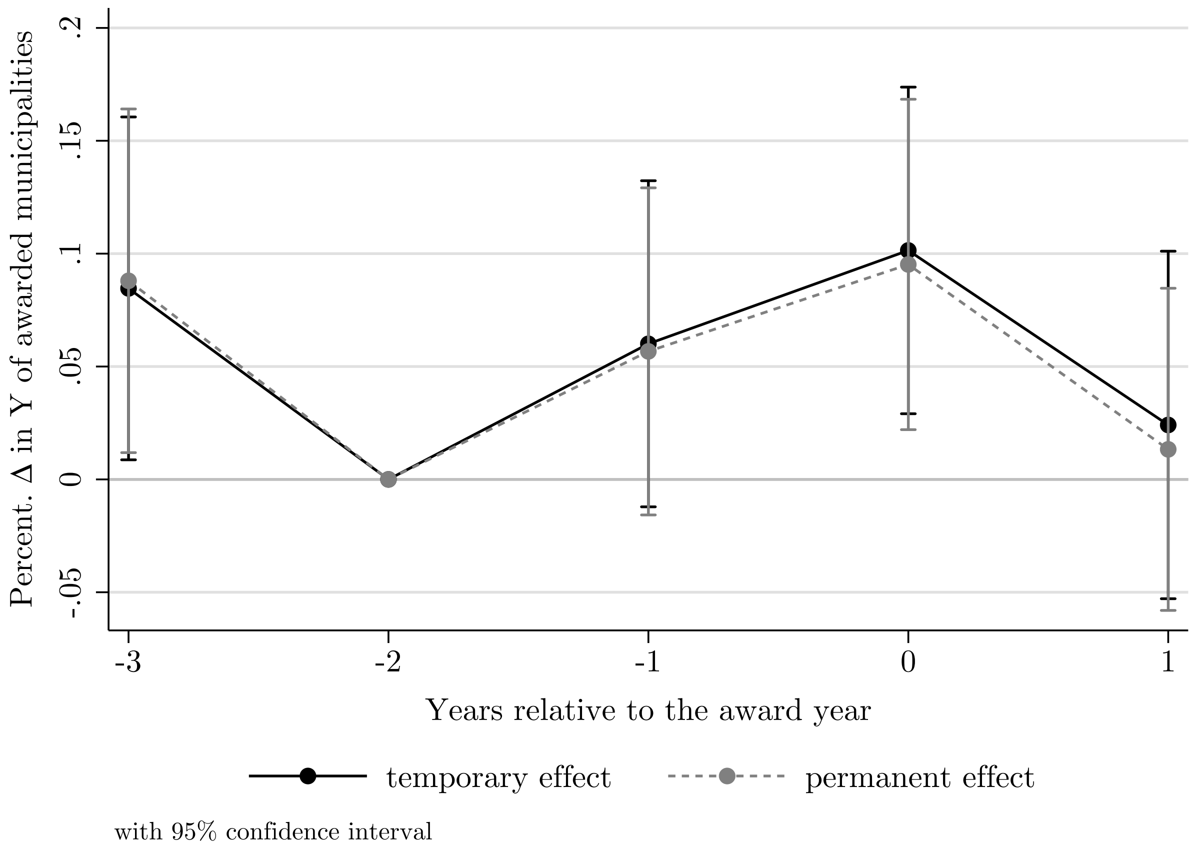


Figure 7: *Temporary and permanent impact of being awarded a Blue Flag on municipal revenues.*

6 Conclusion

Acknowledging the eco-labels' potential in bringing economic and environmental incentives together, a recent literature has developed around them with the aim of testing whether these are effective in practice. The literature evaluating the economic impact of eco-labels has produced mixed results, even when focusing on a single case as the Blue Flag programme. I argue that (i) in absence of detailed data on the applicants, any voluntary certification impact evaluation effort will be sensitive to the econometric assumptions made and (ii) that the Blue Flag literature has wrongly ignored the possibility that the certification might be mainly effecting the first time it is awarded. Therefore, (i) I take an event-study approach, allowing me to avoid the issue of constructing a control group from non-winners and (ii) I focus on the effect of being assigned a Blue Flag for the first time. I find that the effect on municipal revenues of being assigned a Blue Flag for the first time is positive and significant, while I find no evidence that they experience an increase in collective tourist accommodation supply. My findings also provide further evidence (see Creo and Fraboni, 2011; Pencarelli et al., 2016; Cerqua, 2017) not only that the Blue Flag award gives mayors an opportunity

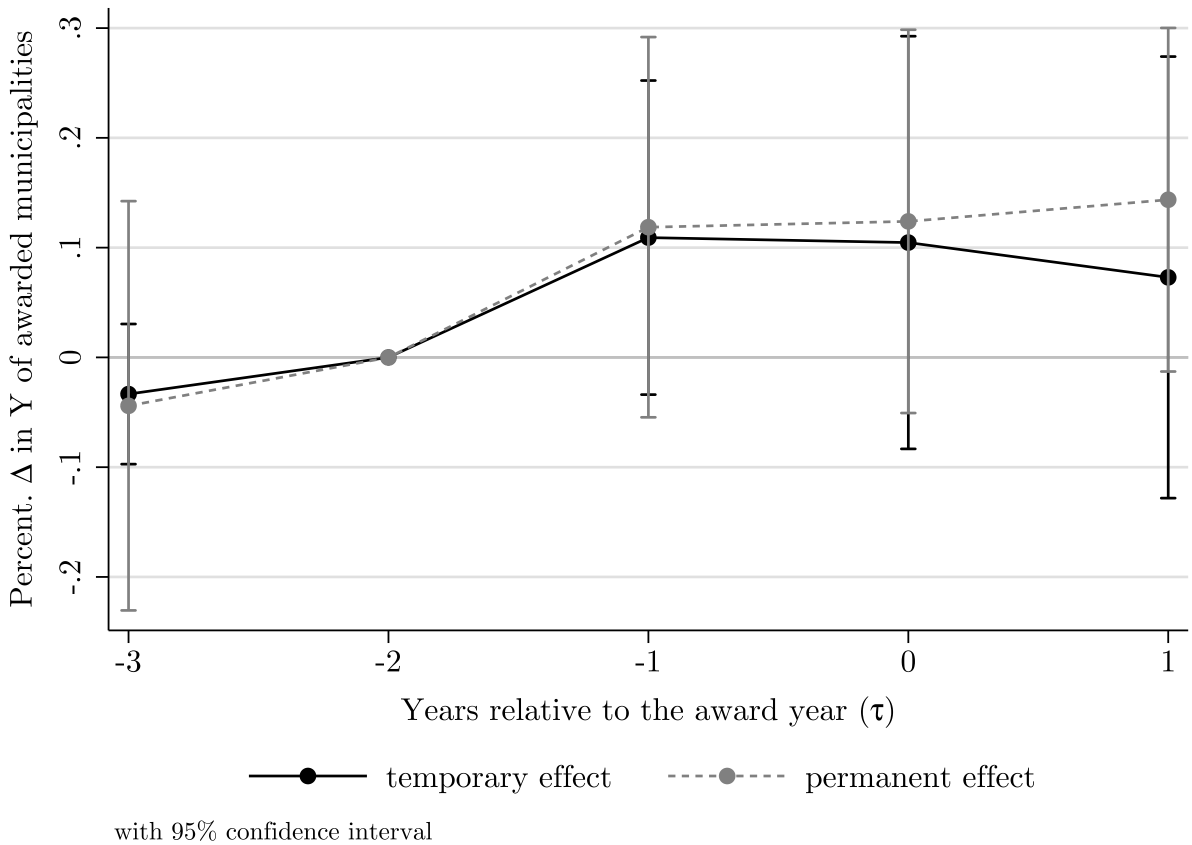


Figure 8: *Temporary and permanent impact of being awarded a Blue Flag on the supply of beds in hotels with 3 stars or more.*

to promote and enact environmentally-conscious infrastructural improvements, but also that they are successful in exploiting it, as evidenced by the increased spending on physical capital.

This is the first study to my best knowledge to provide evidence that municipal revenues are positively affected by a Blue Flag certification. Moreover, if we consider the infrastructural investments required by the Blue Flag as a public good, not a dead-weight cost, then the effect of a fist-time Blue Flag certification on municipal profits is positive (see Pencarelli et al., 2016).

References

Blackman, A., Naranjo, M. A., Robalino, J., Alpízar, F., and Rivera, J. (2014). Does tourism eco-certification pay? costa rica's blue flag program. *World Development*, 58:41 – 52. [Cited on pages 31, 32, 36, and 39.]

Blanco, E., Rey-Maquieira, J., and Lozano, J. (2009). Economic incentives for tourism firms to undertake voluntary environmental management. *Tourism Management*, 30(1):112 – 122. [Cited on page 31.]

Buckley, R. (2002). Tourism ecolabels. *Annals of Tourism Research*, 29(1):183 – 208. [Cited on page 30.]

Capacci, S., Scorcu, A. E., and Vici, L. (2015). Seaside tourism and eco-labels: The economic impact of blue flags. *Tourism Management*, 47:88 – 96. [Cited on pages 31, 32, 34, and 36.]

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The Effect of Minimum Wages on Low-Wage Jobs*. *The Quarterly Journal of Economics*, 134(3):1405–1454. [Cited on pages 32, 35, and 36.]

Cerqua, A. (2017). The signalling effect of eco-labels in modern coastal tourism. *Journal of Sustainable Tourism*, 25(8):1159–1180. [Cited on pages 31, 32, and 43.]

Creo, C. and Fraboni, C. (2011). Awards for the Sustainable Management of Coastal Tourism Destinations: The Example of the Blue Flag Program. *Journal of Coastal Research*, 2011(10061):378 – 381. [Cited on pages 31 and 43.]

FEE (2017). Blue flag programme. [www.blueflag.global; Visited on 25-October-2019]. [Cited on pages 33 and 34.]

Jha, V., Vossenaar, R., and Zarrilli, S. (1997). *Eco-labelling and international trade*. Springer. [Cited on page 30.]

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. *Political Analysis*, pages 1–20. [Cited on pages 32 and 36.]

McKenna, J., Williams, A. T., and Cooper, J. A. G. (2011). Blue flag or red herring: Do beach awards encourage the public to visit beaches? *Tourism Management*, 32(3):576 – 588. [Cited on pages 30 and 31.]

Pencarelli, T., Splendiani, S., and Fraboni, C. (2016). Enhancement of the "blue flag" eco-label in italy: an empirical analysis. *Anatolia*, 27(1):28–37. [Cited on pages 31, 43, and 44.]

Thøgersen, J., Haugaard, P., and Olesen, A. (2010). Consumer responses to ecolabels. *European Journal of Marketing*, 44(11/12):1787–1810. [Cited on page 30.]

Zielinski, S. and Botero, C. (2015). Are eco-labels sustainable? beach certification schemes in latin america and the caribbean. *Journal of Sustainable Tourism*, 23(10):1550–1572. [Cited on page 31.]

Zielinski, S. and Botero, C. M. (2019). Myths, misconceptions and the true value of blue flag. *Ocean & Coastal Management*, 174:15 – 24. [Cited on pages 31, 39, and 40.]