Universität St.Gallen

Behavioral Spillovers

Petyo Bonev

March 2023 Discussion Paper no. 2023-03

School of Economics and Political Science,
Department of Economics

University of St.Gallen

Editor: Mirela Keuschnigg
University of St.Gallen
School of Economics and Political Science
Department of Economics
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen

Electronic Publication: http://www.seps.unisg.ch

Behavioral Spillovers[1]

Petyo Bonev

Author's address:

Petyo Bonev (Adjunct Professor of Econometrics)
Email: Petyo.Bonev@unisg.ch
Swiss Institute for Empirical Economic Research
University of St. Gallen
Varnbüelstrasse 14
9000/St. Gallen
Switzerland

Agroscope,
Tänikon 1, 8356 Ettenhausen
Switzerland
Email: Petyo.Bonev@agroscope.admin.ch

[1] I am greatly thankful to Caterina Alacevich, Jeanine Ammann, and Julien Picard for their helpful comments. I am also thankful to John Thøgersen for his encouraging comments and for bringing the topic into economics.

Abstract

What is a behavioral spillover? How can a spillover be uncovered from the data? What is the precise link between the underlying psychological theory of a spillover and the econometric assumptions which are necessary to estimate it? This paper draws on recent advancements in causal inference, behavioral economics, psychology, and neuroscience to develop a framework for the causal evaluation and interpretation of behavioral spillovers. A novel research design is suggested. The paper challenges existing empirical strategies and reevaluates existing empirical results.

Keywords

Behavioral spillovers, environmental policy evaluation, moral licensing, self-perception theory, cognitive dissonance theory, foot-in-the-door effect

JEL Classification

C21; C26; C9; D04; D9.

1 Introduction

A policy intervention that targets a given individual behavior may lead to unexpected changes in behaviors not targeted by the policy (Ek and Miliute-Plepiene, 2018). Such effects are commonly referred to as behavioral spillovers.[1] Behavioral spillovers may either enhance the policy in the intended direction or reduce the policy impact and even lead to a negative overall effect Ek (2018). Researchers have therefore suggested that policy evaluation should follow an integrated approach which takes explicitly into consideration both targeted and nontargeted behaviors, as well as possible spillovers between them (Thøgersen, 1999). Recently, behavioral spillovers have been extensively studied in multiple domains such as psychology (Truelove et al., 2014; Maki et al., 2019) and economics (Altmann et al., 2022; Bulte et al., 2021; Nafziger, 2020). They have received particular attention in the context of environmental policy evaluation because individual contributions to the common environmental goal can be achieved through multiple activities (Alacevich et al., 2021).

Yet, despite the importance of this topic, the research field has not established a common framework for the *causal* empirical evaluation of behavioral spillovers. One major potential reason is that there are more than one definition of spillovers in the literature (Maki et al., 2019; Galizzi and Whitmarsh, 2019). However, this plurality of definitions is problematic: existing empirical estimates are most commonly interpreted as a spillover without an explicit invoking of one of the definitions, which complicates the comparability and interpretation of the results. This vagueness is often reflected in the scientific language: "a policy might backfire" or "unintended policy effects" could well describe different definitions of behavioral spillovers.

A second potential reason is that behavioral spillovers are predicted by a large number of psychological mechanisms. Notable examples are ego depletion, moral licensing, cognitive dissonance, and the foot-in-the door mechanisms (Dolan and Galizzi, 2015). However, these theoretical explanations do not relate uniformly to the same definition of a behavioral

[1] Jessoe et al. (2021) uses also the term "cross-sectoral" spillovers.

spillover. The subtle differences in the relation between psychological theory and spillover definitions have remained overlooked by the existing literature because research hypotheses are not formally stated as statistical hypotheses and are not explicitly linked to test statistics.

A third reason concerns the objectives of the causal evaluation. While there is a common agreement in the scholar community that knowledge of spillovers is crucial in policy context, existing research has not specified *what this knowledge should be used for in the first place.* In this paper, I show that different scientific objectives require different definitions of spillovers.

This paper makes several contributions to the literature. First, it develops a unified framework for the evaluation and interpretation of behavioral spillovers. The framework is embedded in a formal causal model. Spillover effects are defined using the causal concept of potential outcomes (Imbens and Rubin, 2015). The merit of this formalization is that it makes it straightforward to establish the link between psychological theories and definitions of spillovers over explicitly defined statistical hypotheses and the corresponding test statistics.

The causal framework leads to two unexpected results that challenge existing empirical strategies. First, a randomized experiment is neither sufficient nor necessary to estimate spillover effects. Intuitively, the reason is that adopting a behavior is an individual choice. Policy randomization alone cannot account for the endogeneity of choice. Second, between any two given behaviors, there are in general not only one but many spillovers to be considered - one for each policy regime. Estimating only one spillover, as currently practiced in the literature, amounts to averaging across these multiple spillovers, which leads to a loss of valuable information and hampers the interpretability of the estimates.

The second and major contribution of this paper is to develop a comprehensive set of strategies for the empirical evaluation of behavioral spillovers in the context of policy implementation. The empirical strategies draw on advancements in econometrics, behavioral economics, psychology, and neuroscience. Particular attention is paid to behavioral policies such as green nudges. A novel research design for the evaluation of spillovers is derived. This design combines randomized experiments with surveys that are administered before

the implementation of the policy.

To the best of my knowledge, this is the first paper on behavioral spillovers to follow an integrated multidisciplinary approach. Existing studies have not used a formal causal model (Galizzi and Whitmarsh, 2019; Alacevich et al., 2021; Jones et al., 2019). My paper complements the valuable insights of these studies by strengthening the links between definitions of spillovers, psychological theory, and causal empirical strategies. I also show that the focus of these studies on randomized policies is misguided if the research objective is to evaluate cross-behavioral spillovers and to link them to psychological theory. In contrast, my proposed strategies provide the necessary econometric tools to estimate spillovers and test psychological theory.

This paper also provides an unexpected contribution to the debate on whether (and in which settings) individuals behave rationally in a neoclassical sense. In an influential study, Chetty (2015) argues that this debate should be approached from a pragmatic instrumentalist approach in the tradition of Friedman (1953). Specifically, bounded rationality should be incorporated in a given model of behavior whenever this improves the empirical power of the model. My findings provide a fresh angle on this approach. In particular, the validity of the econometric strategy crucially depends on nontestable assumptions about individual deliberation costs: the more "bounded" the rationality of the individual (in the sense of scarcity of cognitive resources), the easier it becomes to motivate the econometric strategy. Thus, determining the empirical power of a model - which itself depends on the validity of the underlying econometric strategy - requires an ex ante decision in favor of or against a bounded rationality assumption.

A final contribution of my paper is to re-evaluate existing empirical evidence on behavioral spillovers. For tractability reasons, I focus on studies reviewed in the recent meta-study analysis by Maki et al. (2019). I re-evaluate these studies with respect to two criteria. The first one is whether the estimate corresponds to the invoked definition of spillover and to the invoked theories of spillover effects. The second one is whether the study design actually

allows to establish these correspondences. My main finding is that existing empirical results provide no evidence for behavioral spillovers and the psychological theories invoked in the papers. However, I find that several of the evaluated research designs can easily be modified to estimate behavioral spillovers, which lays out a research agenda for future research.

The paper is structured as follows. In section 2, I develop the causal framework and discuss the three existing definitions of a behavioral spillover. Section 3 develops empirical strategies for the evaluation of spillovers. In section 4, existing empirical evidence is re-evaluated.

2 Behavioral spillovers: what they are, what they are not, and why it matters

2.1 Notation and a causal framework

Consider two distinct behaviors of a given individual i, behavior 1 and behavior 2. Let the binary random variable B_{1i} indicate whether individual i adopts behavior 1 ($B_{1i} = 1$) or not (B_{1i}=0). The variable B_{2i} is defined for behavior 2 analogously. For notational simplicity, the individual index i will be dropped when ambiguity is not possible. Let the binary random variable P_i indicate whether individual i is exposed to a given policy intervention ($P_i = 1$) or not ($P_i = 0$). Suppose that this policy has been designed by the policymaker with the objective to incentivize the adoption of behavior 1, while behavior 2 is not considered by the policymaker. In the following, behaviors 1 and 2 will be referred to as the *targeted and nontargeted* behaviors, respectively.

Empirical example 1. Jessoe et al. (2021) study behavioral spillover effects in the context of residential water and electricity consumption. The policy intervention P is a so-called social norm: a home water report that compares a household's water use to that of similar neighbors and provide conservation tips and information about water use. The

targeted behavior is water saving (or reduction of water consumption), while the nontargeted behavior is a reduction in electricity consumption. **Empirical example 2.** Brown et al. (2013) study the effect of setting the temperature in commercial offices (via a thermostat) to a given default value on the actual choice of the temperature by the employees working in these offices. The policy intervention P here is the default option while the targeted behavior B_1 is adopting environmentally friendly heating behavior (e.g. choice of the temperature below 20 degrees Celsius). B_2 is any behavior not targeted by the policy, for example room heating, see Goetz et al. (2022) for a study of spillovers between hot water consumption and heating behavior.

To model behavioral spillovers, I will embed the analysis in the Rubin Causal Framework (Imbens and Rubin, 2015). For any value of the policy $p \in \{0, 1\}$ and any value of behavior $b_1 \in \{0, 1\}$, let $B_{1i}(p), B_{2i}^p(p), B_{2i}(b_1)$ denote the corresponding potential outcomes. As an example, $B_{1i}(1)$ represents the (random) behavioral outcome of individual i in the hypothetical case that she had been exposed to the policy ($p = 1$). The other variables are defined analogously. The superscript in $B_{2i}^p(p)$ is necessary to distinguish between the behavioral outcomes caused by a policy intervention $B_{2i}^p(1)$ and the outcomes caused by a change in the first behavior $B_{2i}(1)$. Put differently, the superscript is necessary to indicate which is the treatment - the policy or the first behavior.

For the sake of simplicity, I focus on additive treatment effects. First, define

$$\delta_i^{target} = B_{1i}(1) - B_{1i}(0)$$

to be the additive treatment effect of the policy on the targeted behavior (behavior 1) for a given individual i. The average treatment effect of the policy on behavior 1 for the whole population can be written as

$$\Delta^{target} = \mathbb{E}[B_{1i}(1) - B_{1i}(0)] = \mathbb{E}[\delta_i^{target}].$$

Similarly, the individual and average treatment effects of the policy on the nontargeted behavior are defined as

$$\delta^{nontarget} = B^p_{2i}(1) - B^p_{2i}(0) \qquad \text{and} \qquad \Delta^{nontarget} = \mathbb{E}[B^p_{2i}(1) - B^p_{2i}(0)] = \mathbb{E}[\delta^{nontarget}],$$

respectively, while the individual and average treatment effects of behavior 1 on behavior 2 are defined as

$$\delta^{spillover} = B_{2i}(1) - B_{2i}(0) \qquad \text{and} \qquad \Delta^{spillover} = \mathbb{E}[B_{2i}(1) - B_{2i}(0)] = \mathbb{E}[\delta^{spillover}].$$

These effects can be visualized with a causal graph, see figure 1. The nodes represent variables (the policy intervention and the two behaviors), while the arrows represent causal relationships. Arrow (1) represents the effect of the policy on the targeted behavior Δ^{target}. Arrow (2) represents the *direct* effect of the policy on the nontargeted behavior *relative* to the targeted behavior, that is, the impact that the policy exerts on behavior through all channels other than through an impact on behavior 1. Arrow (3) represents the effect of adopting behavior 1 on behavior 2 ($\Delta^{spillover}$). The causal path through arrows (1) and (3) represents the indirect effect of the policy on behavior 2. This indirect effect is denoted by $(1) \times (3)$. This notation is justified in certain cases, as I discuss below. The total effect of the policy on behavior 2 ($\Delta^{nontarget}$) is the sum of the two effects (2) and $(1) \times (3)$. The causal graph in figure 1 represents the minimal causal graph for defining behavioral spillovers in the context of a policy intervention.

2.2 Three definitions of behavioral spillovers.

There are three different definitions of behavioral spillovers in the literature (Maki et al., 2019). The first one requires a behavioral spillover to be the effect of adopting one behavior on the second behavior. This definition corresponds to $\delta^{spillover}$ and $\Delta^{spillover}$ (arrow (3)). In the second definition, a spillover is simply the total effect of the policy on the nontar-

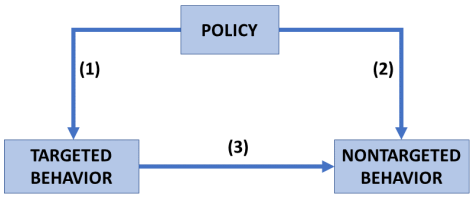

Figure 1: A causal graph for defining behavioral spillovers.

geted behavior. In the causal framework above, this definition corresponds to $\delta^{nontarget}$ and $\Delta^{nontarget}$. The third definition considers a behavioral spillover to be simply the correlation between the two behaviors, $\Delta^{corr} = corr(B_1, B_2)$.

Note that while $\Delta^{nontarget}$ may capture undesired policy consequences, it does not require that there is an effect of the policy on the first behavior, or that there is an effect of behavior 1 on behavior 2. In fact, the definition does not require that the policy targets any behavior at all, i.e. $\Delta^{nontarget}$ is well-defined even without specifying B_1. To estimate $\Delta^{nontarget}$, it suffices to have a randomized policy. Similarly, Δ^{corr} does not require any causal relationship between P, B_1, and B_2. In particular, according to the Reichenbach's principal of causality (Reichenbach, 1988), the correlation between behaviors 1 and 2 may have been induced by a common cause. Δ^{corr} can be estimated from the observed data without any restriction on the data generation process. A more subtle property of Δ^{corr} is that, in contrast to $\Delta^{spillover}$ and $\Delta^{nontarget}$, Δ^{corr} does not require any relation between potential and measured outcomes.

With these considerations in mind, I refer henceforth only to $\Delta^{spillover}$ as behavioral spillovers, while $\Delta^{nontarget}$ and Δ^{corr} are referred to as total policy effect on the nontargeted behavior and (positive or negative) behavioral congruence, respectively.

The relationship between the total policy effect and the behavioral spillover effect: a numerical example. The following numerical example demonstrates that the

total policy effect $\Delta^{nontarget}$ and the behavioral spillover effect $\Delta^{spillover}$ can differ both in sign and magnitude. Assume that P (i) is fully randomized and (ii) has no direct effect on B_2. This is a case of an exclusion restriction which will be discussed in detail in section 3.2. For simplicity, P is assumed to have a constant individual treatment effect on B_1, $\delta^{target} = B_{1i}(1) - B_{1i}(0) = -0.1$ for all i. This implies that the effect on the targeted behavior is $\Delta^{target} = -0.1$. Furthermore, the behavioral spillover is also assumed constant and equal to 0.8,

$$\delta^{spillover} = B_{1i}(1) - B_{1i}(0) = 0.8 = \Delta^{spillover}.$$

The setup is summarized in figure 2. There is no direct arrow from the policy to behavior 2 which reflects the exclusion restriction. The numbers below the arrows represent the two treatment effects.

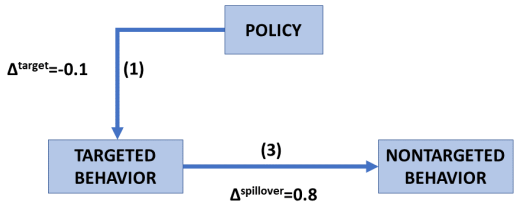

Figure 2: A numerical example with exclusion restriction

With these assumptions, the following lemma can be stated.

Lemma 1 *Suppose that potential and measured outcomes can be linked through the Stable Unit Treatment Assumption.*[2] *Then the total policy effect on B_2 can be written as*

$$\Delta^{nontarget} = \Delta^{target} \times \Delta^{spillover} = 0.1 \times (-0.8) = -0.08. \quad (1)$$

[2]A definition and a detailed discussion of the validity of the Stable Unit Treatment Assumption can be found in section 3.4.

A formal derivation of this result can be found in Appendix A. Importantly, while the total policy effect on B_2 is negative and of very small magnitude (-0.08), the behavioral spillover effect is sizable and positive (0.8). Notably, in a case, in which the policy has a direct effect on B_2, the sign of the behavioral spillover effect cannot be learned directly from Δ^{target} and $\Delta^{nontarget}$. This holds because the direct effect of the policy is also unknown. These considerations can be summarized as follows:

Result 1 *The total policy effect on the nontargeted behavior $\Delta^{nontarget}$ and the behavioral spillover effect $\Delta^{spillover}$ need not agree in sign and/or magnitude. Moreover, whenever P has a direct effect on B_2, $\Delta^{spillover}$ cannot be inferred from $\Delta^{nontarget}$ and Δ^{target} alone.*

In the following, I present two reasons why distinguishing between these definitions matters: a reason related to policy analysis and a reason related to psychological theories of behavior.

Why result 1 matters, reason 1: linking the three definitions to policy analysis. Consider the taxonomy of policy evaluation problems described in the seminal work by Heckman and Vytlacil (2007). Three distinct policy problems are outlined:

(P-1) Evaluating the impact of historical interventions. (P-2) Forecasting the impacts of interventions implemented in one environment in other environments. (P-3) Forecasting the impacts of interventions never historically experienced to various environments.

There is an implicit continuity argument in the distinction between (P-1, P-2) and (P-3). Evaluation of past policies is useful for designing future policies if (i) the new policy and the new environment are not too different from the past ones and (ii) the effect varies smoothly with the policy design and the environment. Under these assumptions, $\Delta^{nontarget}$ provides a good prediction for the effect of a future policy.

On the other hand, if the envisioned policy is very different from past policies, it may be very difficult to guess its effect on the basis of past effect estimates. In such cases, knowledge

of "the behavioral spillovers network", i.e. the spillover effects $\Delta^{spillover}$ between all different pairs of related behaviors, is necessary to assess the total effect of the novel policy. As an example, in 2022, the UK Department for Work and Pensions (DWP) launched an ambitious project based on behavioral nudges aiming to increase saver engagement with the sustainability of pension investments.[3] The environmental aspect of this novel financial behavioral policy implies that nontargeted environmental behaviors may be affected as well. Past estimates of the spillovers $\Delta^{spillover}$ between green pension decisions and different environmental behaviors allow to better predict the overall environmental impact of the policy. In contrast, if past policies targeting pension behavior did not contain an environmental component, the overall effect of the green nudge policy may be impossible to predict based only on past estimates of $\Delta^{nontarget}$.

Why result 1 matters, reason 2: targeting vs. enhancing.[4] A related reason to distinguish between the three definitions is how precisely the knowledge can be used. In particular, knowledge of $\Delta^{spillover}$ can be used to target those behaviors that induce large beneficial spillovers on other behaviors.

Why result 1 matters, reason 3: the link to psychological theory. Suppose that one of the estimates $\hat{\Delta}^{spillover}$, $\hat{\Delta}^{nontarget}$, or $\hat{\Delta}^{corr}$ has been obtained from the data, and assume that this estimate has been correctly attributed to one of the three definitions above. Which psychological theory can be supported or refuted by this estimate? Answering this question is not about determining the sign of the above estimates, but rather about determining which of the estimates $\hat{\Delta}^{spillover}$, $\hat{\Delta}^{nontarget}$, or $\hat{\Delta}^{corr}$ can be used as a test statistics for testing the hypotheses implied by a given theory of spillovers.

The main finding of this section is that existing psychological theories of behavioral

[3]For more information, https://www.gov.uk/government/news/green-nudge-to-greater-knowledge-cop26-trials-launched.

[4]I am thankful to Julien Picard for pointing this reason to me.

spillovers imply one of the statistical hypotheses

$$H_0^+ : \Delta^{spillover} \geq 0 \qquad \text{or} \qquad H_0^- : \Delta^{spillover} \leq 0 \tag{2}$$

To see this, note first that psychological theories of behavioral spillovers can be divided in two broad categories: theories based on common motives and theories based on common resources. According to the former, individual choices and behaviors are driven by deep motives or satisfaction accounts such as maintaining good health, keeping the own environmental footprint low, and others, see Dolan and Galizzi (2015) for an overview of such theories. Any relationships between distinct behaviors are determined by their contributions to these motives. Importantly, all common-motives theories envision a *sequential decision procedure*, in which the individual first decides whether to engage or not in behavior 1, and then, based among others on this decision and its outcome, the individual subsequently decides whether to adopt behavior 2. A necessary component for this type of procedure is some element of either bounded rationality or imperfect information at the first stage of the decision problem. This element allows the individual to learn from engaging in behavior 1. The central example for such a process is an individual who learns about herself from engaging in behavior 1. Such a learning process is suggested by the self-perception theory (Bem, 1972). According to the cognitive dissonance theory, emotions that arise from the perceptions of past behaviors may trigger responses in future behaviors because of the basic individual motive to behave consistently (Festinger, 1957; Thøgersen, 2004). The element of imperfect information or bounded rationality is precisely the implicit assumption that, before the first behavior, the individual does not have full knowledge of herself. Such an assumption is necessary to justify a sequential decision process, as opposed to a simultaneous decision process, in which an individual with perfect knowledge of herself and a perfect foresight of the impact of her behavior decides simultaneously on all future behaviors. In this latter simultaneous decision process, there is no room for behavioral spillovers in a causal

sense.

The common-resources type of theories of behavioral spillovers drops the assumption that behaviors are related through common goals. These theories only require that behaviors compete for scarce cognitive resources. Thus, engaging in one behavior makes it cognitively more difficult to engage in a second behavior, see Nafziger (2020) for a theoretical model of spillover effects caused by limited attention and Altmann et al. (2022) for experimental evidence. Again, there is a causal impact of engaging in behavior 1 on (the intensity of) engaging in behavior 2, and the underlying mechanism is (similarly to the common-motives theories) based on bounded rationality of the individual. In particular, behavior 1 impacts behavior 2 because it (behavior 1) depletes the resources for the subsequent behavior 2.

These considerations lead to the following insight:

Result 2 *Psychological theories of behavioral findings imply one of the statistical hypotheses formulated in* (2)*. Hence, they can be tested using the test statistics* $\hat{\Delta}^{spillover}$*. Together with result 1, this implies that an estimate* $\Delta^{nontarget}$ *of the total policy effect alone does not provide evidence for or against any theory of behavioral spillovers.*

This result is used in section 4 to evaluate the link between existing experimental evidence psychological theories of behavioral spillovers.

Remark: micro-foundations of behavioral spillovers. Similarly to psychological theories, microeconomic models may relate to both $\hat{\Delta}^{spillover}$ and $\Delta^{nontarget}$ or to $\Delta^{nontarget}$ alone. The same informational and cognitive criteria apply here. As an example, Picard (2022) introduces a formal multi-period model of spillovers and policy interventions, in which individuals may be either short-sighted or long-sighted. Short-sighted individuals do not consider future consumption when deciding on the consumption in current periods, so that a decision is made sequentially. Long-sighted individuals, on the contrary, make a consumption-path decision ex-ante for all periods. While this framework allows for $\hat{\Delta}^{spillover}$ in the former case, it precludes behavioral spillovers in the latter.

3 The empirical evaluation of behavioral spillovers

3.1 The endogeneity of the targeted behavior

Consider first $\Delta^{nontarget}$. Since this is a policy effect, there is a large and well-understood econometric toolkit for designing an estimate $\hat{\Delta}^{nontarget}$ that satisfies at least one of the two properties

$$\mathbb{E}[\hat{\Delta}^{nontarget}] = \Delta^{nontarget} \text{ (no bias)}; \qquad \hat{\Delta}^{nontarget} \xrightarrow{p} \Delta^{nontarget} \text{ (consistency)}. \tag{3}$$

The most common strategy in the empirical literature is to randomize the policy P. In example 1, Jessoe et al. (2021) implement a framed field experiment, in which randomly chosen households receive the water report ($P = 1$), while all other households do not receive it ($P = 0$). With a randomized P, $\hat{\Delta}^{nontarget}$ can be constructed as a difference in means of the outcomes for treated and nontreated, that is,

$$\hat{\Delta}^{nontarget}_{RCT} := \hat{\mathbb{E}}[B_2|P = 1] - \hat{\mathbb{E}}[B_2|P = 0] := \frac{1}{n_1} \sum_{i:P_i=1} B_{2i} - \frac{1}{n_0} \sum_{i:P_i=0} B_{2i}, \tag{4}$$

where n_1 and n_0 are the numbers of treated and nontreated individuals in the sample, respectively. This is the standard estimator in the literature and it satisfies both properties (3).

Estimating $\Delta^{spillover}$, on the other hand, is not straightforward at all. First, note that result 1 implies that any estimator $\hat{\Delta}^{nontarget}$ that has the properties (3) is neither a consistent nor an unbiased estimator for $\Delta^{spillover}$. Notably, this holds also for the standard policy estimator $\hat{\Delta}^{nontarget}_{RCT}$ defined in (4). The following result summarizes these insights:

Result 3 *The standard estimator $\Delta^{nontarget}_{RCT}$ produced by a randomized controlled trial is neither consistent nor unbiased estimator for the behavioral spillover effect $\Delta^{spillover}$.*

Before studying consistent estimators for $\Delta^{spillover}$, note that $\Delta^{spillover}$ and $\Delta^{nontarget}$ differ

not only in terms of their relation to psychological theories, but also in terms of the sources of variation of P and B_1. While the variation in P is generated by a controlled decision of the policy maker, the variable B_1 (under types 1 and 2 policies) is a choice variable: the decision to adopt behavior 1 is taken by the individual. Moreover, this decision will in general depend on characteristics of the individual.

Empirical example 1, continued. Let B_1 indicate pro-environmental water consumption as in Jessoe et al. (2021). A large and growing number of papers have studied the determinants of pro-environmental behaviors. Psychological variables positively associated with pro-environmental behaviors include willingness to sacrifice, perceived behavioral control, subjective norms, and green self-identity (Hansmann et al., 2020). Other types of relevant factors include socio-economic variables (Blankenberg and Alhusen, 2018), demographic factors such as family structure (Singha et al., 2023), political attitudes (Korfiatis et al., 2004), as well as behavioral factors such as habits (Singha et al., 2023). Importantly, the literature has found that some of these factors are related to the perceived benefits of the environmental behaviors (Hansmann et al., 2020).

The example above highlights two aspects of behavior adoption. First, a rich source of individual characteristics explains pro-environmental behaviors. Second, the relationship between these factors and perceived benefits of behaviors implies a selection based on potential outcomes. Thus, if some of the individual factors are not controlled for (e.g., because they are not observed by the researcher), the estimate of $\Delta^{spillover}$ might be biased. As a result, any econometric strategy for the estimation of the behavioral spillover must take the endogenous choice into account.

Developing a strategy for dealing with the endogeneity of B_1 requires to distinguish between two different cases. In the first one, the policy has no direct effect on B_2. This corresponds to the causal graph in figure 2. The lack of a direct arrow from the policy to the nontargeted behavior corresponds to the lack of a direct effect. In the second case, a direct effect is possible. This is the general case and corresponds to figure 1.

3.2 Evaluation of $\Delta^{spillover}$ under the assumption of no direct effect

3.2.1 A nonparametric estimation approach

Consider first the case in which the policy has no direct effect on behavior 2. This assumption is referred to as an *exclusion restriction* in the econometric literature (Heckman and Navarro-Lozano, 2004). When the exclusion restriction holds and P is exogenous, P can be used as an instrumental variable for the endogenous B_1.

To ensure generality of the instrumental approach, an approach is needed that imposes no parametric assumptions and no restrictions on the heterogeneity of the treatment effect. Estimates from parametric instrumental variable models such as the Two Stage Least Squares are difficult to link to treatment effects (Crudu et al., 2022; Goodman-Bacon, 2021). Similarly, the assumption of constant treatment effects in the numerical example above is unrealistic and needs to be relaxed.

An novel insight of this paper is that a standard econometric approach referred to as Local Average Treatment Effect (LATE) estimation approach can be adapted to evaluate behavioral spillovers. The LATE estimation approach can be used to estimate $\Delta^{spillover}$ under four major assumptions (Imbens and Angrist, 1994). A1: P is randomized (exogeneity assumption); A2: P has no direct impact on the second behavior (exclusion restriction); A3: a monotone effect of P on B_1 (monotonicity) A4: the Stable Unit Treatment Assumption (SUTVA). While the exogeneity assumption is trivially satisfied in a randomized controlled experiment, the other three assumptions require a careful justification and are discussed below.[5]

Under the above mentioned assumptions, the LATE approach allows to estimate the behavioral spillover effect for the subgroup of compliers. This group consist of all individuals that are affected by the policy, i.e. they change their behavior 1 as a result of the policy intervention. For these individuals, the spillover effect can be estimated from the following

[5]Since SUTVA is essential for both strategies (with and without a direct effect), it is discussed after the second strategy is developed.

equation:

$$\Delta^{spillover}_{compliers} = \frac{\Delta^{nontarget}}{\Delta^{target}}. \tag{5}$$

Unlike result (1), result (5) pins down only the effect for the compliers. In contrast, individuals who either always adopt behavior 1 (referred to as "always takers"), or who never adopt it ("never takers") are not affected by the policy. Therefore, random variation in P is not informative for these individuals and their spillover effects cannot be estimated.

3.2.2 The problematic justification of the exclusion restriction

The exclusion restriction is a non-testable assumption. Thus, its validity can only be discussed in the context of existing psychological theories. I clarify this point in the following example.

Empirical example 2, continued. Does a thermostat default(P) affect directly a nontargeted behavior B_2 such as water consumption? The answer depends on the precise mechanism of how the policy affects the *targeted* behavior. It can be broadly distinguished between two types of mechanisms.

The first one is a scarce cognitive resources mechanism. Individuals do not change the thermostat temperature because of the associated cognitive cost, a mechanism referred to as inertia in the literature (Brown et al., 2013; Hedlin and Sunstein, 2016). The nudge helps individuals not to think about the targeted behavior. More importantly, this channel leaves the motives of an individual unaffected. Behavior 1 is executed/adopted "automatically" because it is the straightforward behavior to adopt (behavior associated with least cognitive cost). It is therefore plausible to assume that the policy does not have a direct effect on other environmental behaviors simply because the policy is ignored by the individual.

The second one is an information mechanism. The default may be interpreted by the individual as an implicit recommendation by the policy maker, see McKenzie et al. (2006)

for convincing empirical evidence. Implicit recommendations of policies play a central role in the literature on crowding of intrinsic motivation. When an individual interprets a policy as an implicit recommendation, this recommendation might as well be used by the individual to infer the intentions of the policy maker (Bowles and Polania-Reyes, 2012). As a result, the motivation of the individual to act pro-socially might be affected positively (crowding in) or negatively (crowding out). Although this channel has been discussed in the context of a single behavior, there is no reason to assume that crowding of intrinsic motivation would leave other behaviors unaffected. In the context of behaviors connected through common motives, this implies that the policy would have a direct effect on the nontargeted behavior.

The above examples lead to the following important result:

Result 4 *The validity of the exclusion restriction depends on the psychological mechanism of how a policy affects behavior 1. Bounded rationality theories with cognitive scarcity may be compatible with an exclusion restriction, while information mechanisms are not.*

Importantly, since there is no general empirical evidence that precludes the second type of explanations, the exclusion restriction is in practice very hard to defend.

3.2.3 The monotonicity assumption and reactance

The LATE estimation approach requires that the effect of P on the targeted behavior is monotonic. Using potential outcomes notation, the monotonicity assumption can be stated as

$$B_1(0) \leq B_1(1). \tag{6}$$

The following example gives the intuition behind (6).

Empirical example 2, continued. In the context of the thermostat nudge of Brown et al. (2013), monotonicity (6) amounts to assuming that there are no individuals who would behave conversely to the policy - that is, individuals who would behave pro-environmentally

($B = 1$) if there is no default ($P = 0$) but who would increase the temperature above 20 degrees ($B = 0$) if they are subject to the default ($P = 1$).

Because of their counter-policy behavior, such individuals are also referred to as "defiers" (Imbens and Angrist, 1994). The monotonicity requirement is thus equivalent to the assumption that there are no defiers.

But is this seemingly irrational behavior indeed unlikely? Psychologists and neuroscientists have recently paid increasing attention to a phenomenon referred to as reactance (Brehm and Brehm, 2013). Reactance is "an unpleasant motivational arousal that emerges when people experience a threat to or loss of their free behaviors" (Steindl et al., 2015). In a policy context, reactance occurs when individuals interpret a given policy as an attempt by the policy maker to control or restrict their choices.

The presence of reactance potentially violates assumption (6). In particular, reactance manifests itself in defiant behavior which aims - consciously or unconsciously - at restoring the autonomy of the individual. Such defiant behavior has been documented in numerous studies on crowding out intrinsic motivation (Bowles and Polania-Reyes, 2012), however, the implications for the econometric evaluation of policies have remained unrecognized.

Result 5 *Crowding out intrinsic motivation because of reactance invalidates the LATE approach because it violates the monotonicity assumption.*

This result is particularly problematic in the context of green nudges because of their perceived manipulative nature. As an example, recent research on green electricity defaults has documented that "when a socially desirable good or service is offered, ...automatic enrollment may backfire as a result of reactance" (Hedlin and Sunstein, 2016). Hence, reactance to green nudges may not only pose threat to the effect of the policy, but also to its evaluation by the econometrician.

3.3 Evaluation of $\Delta^{spillover}$ when a direct effect is possible

3.3.1 Behavioral spillovers under policy interaction

Suppose now that an exclusion restriction cannot be adopted, so that the policy P has a direct effect on B_2. Two complications arise from this direct effect. First, P is not a valid instrument for B_1, and therefore the nonparametric LATE approach cannot be used to estimate $\Delta^{spillover}$. Second, there are now more than one behavioral spillovers to be estimated. To see this, note that under a direct effect of P on B_2, the averages of the individual treatment effects will vary with the policy arm:

$$\mathbb{E}[\delta^{spillover}|P=1] \neq \mathbb{E}[\delta^{spillover}|P=0]. \tag{7}$$

Result (7) is formally shown in appendix B. Intuitively, the policy now modifies the environment, in which the causal effect of behavior 1 on behavior 2 is generated.[6] The parameter $\Delta^{spillover}$ is a weighted average of these two different averages:

$$\Delta^{spillover} = \mathbb{E}[\delta^{spillover}] \tag{8}$$

$$= \mathbb{E}[\delta^{spillover}|P=1]Prob\{P=1\} + \mathbb{E}[\delta^{spillover}|P=0]Prob\{P=0\}. \tag{9}$$

Because an individual is either subject to a given policy or not, $\Delta^{spillover}$ is of no interest to the researcher; it is not informative on either of the two possible states $P=0,1$. Instead, the researcher is interested in the state-specific spillovers.

Result 6 *When P has a direct effect on B_2, there are two behavioral spillovers, one for each treatment arm $P=0,1$.*

Henceforth, I denote these two spillovers by $\Delta_0^{spillover}$ and $\Delta_1^{spillover}$, respectively. Note that each of these two spillover effects has to be estimated in an environment with no variation in P: P is set to either 0 or 1. Thus, a different source of exogenous variation is needed

[6] Using the linear regression jargon, "P interacts with B_1" when generating B_2.

for estimation. There are two types of cases to be considered: the so-called "selection on observables" and "selection on unobservables" settings. A novel result discussed below is that standard econometric approaches for the latter case such as Regression Discontinuity Design cannot be applied to estimate behavioral spillovers.

3.3.2 Selection on observables.

The selection on observables case amounts to assuming that conditionally on the observed individual characteristics X, the treatment assignment of B_1 is ignorable:

$$B_1 \perp\!\!\!\perp (B_2(1), B_2(0))|X; \tag{10}$$

Condition (10) is also referred to as "Conditional Independence Assumption" (CIA) (Lechner, 2001). Intuitively, within any group of individuals sharing the same values of all observed covariates, assignment of B_1 can be treated as generated by a randomized experiment. Importantly, since B_1 represents the choice to adopt behavior 1, the random vector X must contain all individual characteristics that explain that choice and are simultaneously factors of the outcome variable B_2. In the context of B_1 being an environmental behavior, it follows from the discussion of example 1 above that X must include demographic and socio-economic characteristics such as age, number of children, occupation, income, but also environmental and political preferences, as well as cognitive and noncognitive characteristics of the individual.

When the CIA is satisfied for each treatment arm $P = 0, 1$, $\Delta_1^{spillover}$ and $\Delta_2^{spillover}$ can be estimated with any econometric approach that relies on balancing X.[7]

An integrated research design. Most commonly, preferences and psychological characteristics of individuals are not observed in administrative datasets. One way to obtain access to these variables is through a survey.[8] This suggests the following integrated research

[7]Examples for such an approach are the matching, (augmented) inverse probability weighting and doubly robust machine learning estimation approaches.

[8]Importantly, the survey must be administered *prior* to the realization of B_1, B_2. If B_1 is realized prior

design for evaluating behavioral spillovers and policy effects: Step 1 (policy assignment): randomly determine which individuals receive the intervention ($P = 1$) and which will not; Step 2 (survey step): administer a survey that measures a large battery of characteristics X for each treatment arm; Step 3 (estimation): esimate $\Delta^{nontarget}$ using the random assignment of P; estimate $\Delta_1^{spillover}$ and $\Delta_2^{spillover}$ using the observed characteristics X_1 and a balancing approach.

3.3.3 Selection on unobservables.

Suppose now that condition (10) is not satisfied. This is the case when the researcher does not observe sufficiently many covariates to ensure the conditional ignorability of the choice B_1. The econometric literature has developed a variety of strategies that deal with this case. These strategies, commonly referred to as "selection on unobservables" identification methods, rely on exogenous variation of external to the individual conditions or incentives. A standard example is the Regression Discontinuity Design which relies on randomness generated by a threshold in some external observed factor (referred to as "forcing variable").[9]

Crucial feature of such quasi-experimental variation in incentives is that it is always external to the individual. That is, the data generation process that triggers the variation in adopting B_1 is not driven by the individual decision but originates "outside" of the individual. In contrast, individual-driven thresholds or other types of quasi-experimental variation such as intrinsic uncertainty cannot be observed by the econometrician in the selection-on-unobservables setting. Thus, equivalently to the distinction between the policy P and the actual behavior B_1, any source of external variation must be distinguished from the behavior of the individual.

This conclusion brings us back to the initial position of asking the question whether the external source of variation in incentives has a direct effect on B_2 or not. If it has no

to measuring X, then the CIA (10) is violated. The reason is that in these cases, some components of X such as environmental preferences might be affected by the variable they should explain (B_1).

[9]Further examples are the Synthetic Control Approach, the Bunching Approach, and the Difference-in-Differences approach.

direct effect, then an instrument based on this variation and on the exclusion restriction assumption can be used to identify $\Delta_0^{spillover}$ and $\Delta_1^{spillover}$. If an exclusion restriction is not plausible, then the two spillover effects cannot be identified and estimated. In this sense, identification using an exclusion restriction as in section 3.2 and selection on unobservables share an identical requirement. This finding can be summarized as follows.

Result 7 *Behavioral spillovers can either be estimated either using a selection-on-observables strategy or using a strategy based on an exclusion restriction. The latter must be imposed either on P or on the source of quasi-experimnetal variation in a setting with selection on unobservables.*

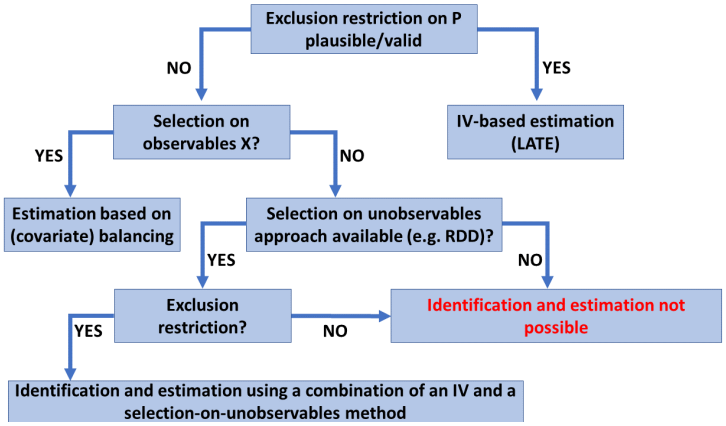

Figure 3: A summary of identification cases

Figure 3 summarizes the different cases and the corresponding identification and estimation approaches. When an exclusion restriction on P is plausible, a nonparametric instrumental variable approach can be used. When the exclusion restriction is violated but the CIA is satisfied, the resulting two spillover effects can be estimated either with a balancing approach. If the CIA is violated as well and a selection-on-unobservables approach for an external incentive is plausible, this approach can be used if the external incentive has no direct effect on B_2.

3.4 The SUTVA assumption: extensive vs. intensive spillover effects

The Stable Unit Treatment Value Assumption requires that requires that there are no equilibrium effects and that there are no different versions of the same treatment (Imbens and Rubin, 2010; Rubin, 1986). SUTVA provides the link between the potential outcomes of an individual and the corresponding actually observed outcomes (B_{1i}. Thus, it underlies all of the above empirical strategies for estimating spillover effects.

I focus on the second condition of SUTVA, which requires that the treatment is the same across units. To that end, it is necessary to distinguish between behavioral spillovers along the extensive and the intensive margins. In particular, thus far the random variable B_1 representing the targeted behavior 1 was modeled as a binary adopt/not adopt variable, $B_1 \in \{0, 1\}$, (the extensive margin). However, the strength of the behavioral spillover may depend on a continuous measure of behavior 1 (the intensive margin). This measure might represent the duration of maintaining a certain behavior. As an example, adopting a vegetarian diet for environmental reasons is likely to have only a marginal positive (or even negative) impact on our self-image if we break the diet after a week compared to permanently maintaining the diet. Alternatively, the behavioral measure might capture the intensity or extent to which we engage in behavior 1. As an example consider the size of a donation (in percentage of the monthly income) to a pro-environmental cause. Both the duration and the intensity cases give rise to a setting, in which a marginal treatment effect of a continuous variable has to be estimated. The effect is marginal because it potentially depends on the baseline level of behavior 1.

Estimating the effect of a the categorical treatment (adop vs. do not adopt a given behavior) when the actual treatment is continuous represents again a violation of the SUTVA behavior. In particular, it violates the "no versions of the same treatment" requirement, since each unit that has adopted a given behavior will typically exert it in a different intensity.

This violation can lead in extreme cases to meaningless estimates.[10]

Potential solutions to multiple versions of the treatment are discussed e.g. in VanderWeele and Hernan (2013) but also in the dose-response literature (Imbens, 2000). Importantly, while the identification requirements represented in figure 3 remain largely the same, estimation with a continuous treatment requires additional assumptions because of the increased dimensionality of the treatment.

3.5 Evaluation of spillover effects when the policy is commanded

Finally, consider the case in which the policy is commanded. An example for such case is when the policy is an implementation of a ban of a given behavior, e.g., a ban of using residential water for gardening. For simplicity, assume that noncompliance is not possible or its cost to the individual is too high. Thus, even though the policy and the given behavior are still different distinct actions, the values of P and B_1 coincide in the sample. As a result, the effects of the treatments P, B_1, and (P, B_1) on B_2 cannot be empirically distinguished, i.e. neither $\Delta^{spillover}$ nor $\Delta^{nontarget}$ can be identified and estimated.

Remark. In the typical setting of lab experiments, experimental subjects are *asked* by the experimenter to perform a given behavior. Here, noncompliance is in fact possible and potentially provides valuable information. The intervention (asking individuals to adopt a behavior) represents a so-called "Intention-to-Treat" variable, which, under the requirements discussed above, can be used as an instrument for the actual targeted behavior. Whenever noncompliance is not observed, the nonidentification of $\Delta^{spillover}$ and $\Delta^{nontarget}$ applies.

4 Re-evaluation of existing evidence

In this section, I re-evaluate existing evidence on behavioral spillovers. For tractability reasons, I focus on the papers evaluated in the meta-study by Maki et al. (2019). From the

[10] This is succinctly demonstrated in the paper "Does water kill?" by Hernán (2016).

25 papers evaluated, I re-evaluate only peer-reviewed publications. In addition, I exclude the paper by Raimi et al. (2019) since both B_1 and B_2 represent attitudes (such as worry about climate change) and not actual behaviors or intentions for behaviors. Finally, I do not list and discuss separately multiple experiments within a given study because these experiments are equivalent in terms of empirical strategy.

The results of my meta-analysis are presented in table table 1. For each paper, I study the following aspects. First, I identify the empirical strategy of that paper. This strategy is recorded in column 1 of table 1. The large majority of papers use a randomized controlled trial (RCT), in which the policy is randomized. The two exceptions are the studies by Poortinga et al. (2013) and Schultz et al. (2015) which use a type of a difference-in-differences and geographical matching strategies, respectively. Next, I record how the studies interpret their main estimates, see column 2 of table 1. In particular, I evaluate whether a given study interprets its estimate as an estimate of a spillover $\Delta^{spillover}$ (in which case the corresponding entry of the table is recorded as $\hat{\Delta}^{spillover}$) or as an estimate $\hat{\Delta}^{nontarget}$ of the total effect of the policy on nontargeted behaviors $\Delta^{nontarget}$. To record this information, I evaluate either the verbally stated research hypotheses (whenever such are explicitly formulated) or simply infer the definition from the general discussion on the effect of interest. In some cases, there is discrepancy between the two sources of information. To infer what effect the authors are really after, I evaluate the discussion on psychological mechanisms behind their theory, see column 3 of table 1. When this information is not available, I simply state that it is unclear what effect is estimated. In column 4, I record what the actual object of estimation is, i.e., which object ($\Delta^{spillover}$ or $\Delta^{nontarget}$) is consistently estimated by the estimator of the study. In a next step, I evaluate whether an exclusion restriction is satisfied (column 5) or the CIA assumption is satisfied (column 6) - the only two strategies that can be used to estimate the actual behavioral spillover effect $\Delta^{spillover}$. Finally, using the information from columns 5 and 6, I evaluate whether the study could have estimated both $\Delta^{spillover}$ or $\Delta^{nontarget}$ with the existing study design. For each evaluated paper, I provide a brief discussion that motivates

my conclusions, see the discussion in appendix C.

Table 1: Reevaluation of studies surveyd in Maki et al. (2019)

Paper	Used strategy	Interpretation of estimate	Invoked psychological explanation	Actual estimate	Exclusion restriction	CIA	Potentially estimable with empirical design
Carrico et al. (2018)	RCT, randomized P	$\hat{\Delta}^{spillover}$	multiple theories	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Geng et al. (2016)	RCT, randomized P	$\hat{\Delta}^{spillover}$	moral licensing	$\hat{\Delta}^{nontarget}$	plausible	not satisfied	$\Delta_1^{spillover}, \Delta_2^{spillover}$, $\Delta^{nontarget}$
Lacasse (2015)	RCT, randomized P	$\hat{\Delta}^{spillover}$	self-perception theory	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Lacasse (2016)	RCT, randomized P	$\hat{\Delta}^{spillover}$	self-perception theory & moral licensing	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Lacasse (2019)	RCT, randomized P	$\hat{\Delta}^{spillover}$	norm-values & self-perception theory	$\hat{\Delta}^{nontarget}$	not satisfied	plausible	$\Delta_1^{spillover}, \Delta_2^{spillover}$, $\Delta^{nontarget}$
Margetts and Kashima (2017)	RCT, randomized P	$\hat{\Delta}^{spillover}$	resource constraints within the goal theory	$\hat{\Delta}^{nontarget}$	not satisfied	plausible	$\Delta_1^{spillover}, \Delta_2^{spillover}$, $\Delta^{nontarget}$
Parag et al. (2011)	RCT, randomized P	$\hat{\Delta}^{spillover}$	mental accounting theory	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Poortinga et al. (2013)	Difference-in, differences	$\hat{\Delta}^{spillover}$	self-perception theory & consistency theory	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Schultz et al. (2015)	geographic matching	unclear	no theory	$\hat{\Delta}^{nontarget}$	plausible	not satisfied	$\Delta_1^{spillover}, \Delta_2^{spillover}$, $\Delta^{nontarget}$
Steinhorst et al. (2015)	RCT, randomized P	$\hat{\Delta}^{nontarget}$	norms activation, goal theory	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Thomas et al. (2016)	Difference-in, differences	$\hat{\Delta}^{spillover}$	multiple	$\hat{\Delta}^{corr}$, $\Delta^{notarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Tiefenbeck et al. (2013)	RCT, randomized P	$\hat{\Delta}^{nontarget}$	moral licensing	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Steinhorst et al. (2015)	RCT, randomized P	$\hat{\Delta}^{nontarget}$	norms activation, goal theory	$\hat{\Delta}^{nontarget}$	not satisfied	not satisfied	$\Delta^{nontarget}$
Truelove et al. (2016)	RCT, randomized P	$\hat{\Delta}^{nontarget}$	guilt, worry, identity	$\hat{\Delta}^{nontarget}$	plausible	not satisfied	$\Delta_1^{spillover}, \Delta_2^{spillover}$, $\Delta^{nontarget}$

Re-evaluation of existing evidence. For each paper, I evaluate the claimed estimate (column 2), the empirical strategy (column 1), the theoretical interpretation (column 3), the actual estimate (column 4), validity of the exclusion restriction and the CIA (columns 5 and 6), and what could be potentially learned from the empirical design (column 7).

Summary of the re-evaluation. Based on the results in table 1, the following conclusions can be made. First, all papers except for Schultz et al. (2015) consider $\Delta^{spillover}$ as the main object of their study and interpret their estimate accordingly as an estimate of $\Delta^{spillover}$.[11] The majority of these papers embed their analysis in the context of the self-perception theory. Second, with the exception of Thomas et al. (2016), all of the surveyed studies actually estimate $\Delta^{nontarget}$.[12] These estimates compare nontargeted outcome averages of treated and nontreated individuals, which correspond to $\hat{\Delta}_{RCT}^{nontarget}$. From these considerations, the following result follows.

Result 8 *The estimates in studies surveyed by Maki et al. (2019) can be used to assess the policy effects on nontargeted behaviors $\Delta^{nontarget}$ but not behavioral spillovers $\Delta^{spillover}$. Accordingly, these estimates cannot be used to evaluate (i.e. either support or refute) the psychological theories these studies invoke.*

In the majority of studied policies, the estimated policy effects are compatible with one of the following explanations. First, the treatment includes a clue that a pro-environmental behavior is desirable (desirability bias). As an example, Carrico et al. (2018) ask participants to reduce red meat consumption in order to reduce emissions of greenhouse gases. Second, the treatment increases the salience of the environmental importance of individual behaviors, see e.g. the study of Lacasse (2015). Third, the treatment crowds out the intrinsic motivation of the individual to behave pro-socially or pro-environmentally. As an example, people who receive a water consumption report as in Tiefenbeck et al. (2013) might feel monitored or controlled and respond with defiance (reactance). None of these three explanations invokes behavioral spillovers - they all describe a direct effect of the treatment on the nontargeted behaviors.

Yet, on a positive note, 5 out of the 14 evaluated studies have a research design that *allows* to estimate behavioral spillovers. In particular, the studies of Margetts and Kashima (2017) and Lacasse (2019) measure a large number of individual covariates through a survey. In these two cases, the CIA may be plausible and the selection on observables estimation approach can be applied. Furthermore, the studies of Geng et al. (2016), Schultz et al. (2015), and Truelove et al. (2016) design a policy whose environmental aspect is hard to infer and which is unlikely to crowd out intrinsic motivation. For these policies, the exclusion restriction appears to be plausible, so that the LATE approach could in principle be used to uncover $\Delta^{spillover}$.[13] These studies give hope for future empirical evaluations of behavioral spillovers.

[11] Schultz et al. (2015) formulate no clear reference to a definition of a spillover.

[12] The study of Thomas et al. (2016) is the only study to regress changes in B_2 on B_1. Yet, since the time sequence of B_1 and B_2 cannot be established, their estimate is potentially biased due to reverse causality and accordingly can be interpreted only as $\hat{\Delta}^{corr}$.

[13] Unfortunately, in two of these three studies, non-compliers - individuals who did not adopt the desired targeted behavior - are thrown out from the sample, so that the variation necessary to estimate a LATE is artificially removed.

A Proof of lemma 1

The result can be derived using the following steps:

1. The total policy effect on B_2 is equal to

$$\Delta^{nontarget} = \mathbb{E}[B_{2i}^p(1) - B_{2i}^p(0)]. \tag{11}$$

Because of the randomization of P, it holds $(B_{2i}^p(1), B_{2i}^p(0)) \perp P$, and hence, using SUTVA, we obtain

$$\Delta^{nontarget} = \mathbb{E}[B_{2i}|P_i = 1] - \mathbb{E}[B_{2i}|P_i = 0]. \tag{12}$$

2. For $\mathbb{E}[B_2|P = 1]$, it holds (i is omitted for simplicity)

$$\begin{aligned}
&\mathbb{E}[B_2|P=1] \overset{B_2 \text{ binary}}{=} Prob\{B_2 = 1|P = 1\} = Prob\{\{B_2 = 1 \cap B_1 = 1\} \cup \{B_2 = 1 \cap B_1 = 0\}|P = 1\} \\
&= Prob\{B_2 = 1, B_1 = 1|P = 1\} + Prob\{B_2 = 1, B_1 = 0|P = 1\} \\
&= Prob\{B_2 = 1|B_1 = 1, P = 1\}Prob\{B_1 = 1|P = 1\} \\
&+Prob\{B_2 = 1|B_1 = 0, P = 1\}Prob\{B_1 = 0|P = 1\} \\
&\overset{\text{excl. restr.}}{=} Prob\{B_2 = 1|B_1 = 1\}Prob\{B_1 = 1|P = 1\} + Prob\{B_2 = 1|B_1 = 0\}Prob\{B_1 = 0|P = 1\} \\
&= \mathbb{E}[B_2|B_1 = 1]\mathbb{E}[B_1|P = 1] + \mathbb{E}[B_2|B_1 = 0](1 - \mathbb{E}[B_1|P = 1]) \\
&= \mathbb{E}[B_2(1)]\mathbb{E}[B_1(1)] + \mathbb{E}[B_2(0)](1 - \mathbb{E}[B_1(1)])
\end{aligned}$$

3. Following analogous steps, we obtain

$$\mathbb{E}[B_2|P = 0] = \mathbb{E}[B_2(1)]\mathbb{E}[B_1(0)] + \mathbb{E}[B_2(0)](1 - \mathbb{E}[B_1(0)]).$$

4. Subtracting $\mathbb{E}[B_2|P = 0]$ from $\mathbb{E}[B_2|P = 1]$, we thus obtain

$$\begin{aligned}
&\Delta^{nontarget} = \mathbb{E}[B_2(1)](\mathbb{E}[B_1(1)] - \mathbb{E}[B_1(0)]) - \mathbb{E}[B_2(0)](\mathbb{E}[B_1(1)] - \mathbb{E}[B_1(0)]) \\
&= \Delta^{targer} \times \Delta^{spillover}.
\end{aligned}$$

B Proof of result (7)

To show (7), note first that when the policy P has a direct effect on behavior 2, the potential outcomes notation has to be adjusted as follows: each potential outcome for behavior 2 has now two arguments, one for the policy intervention and one for the intervention caused by behavior 1. Formally, a potential outcome now is written as

$$B_2(p, b_1) \qquad \text{with } p \in \{0,1\}, b_1 \in \{0,1\}. \tag{13}$$

An exclusion restriction (no direct effect) eliminates the first argument, which leads to the notation used throughout the paper. Without an exclusion restriction, it is necessary to distinguish between the potential outcomes $B_2(1, b_1)$ and $B_2(0, b_1)$. In particular, it must not hold $B_2(1, b_1) = B_2(0, b_1)$. As a result, unless a restrictive assumption on the data generation process is adopted, it holds

$$\mathbb{E}[B_2(1,1) - B_2(1,0)|P = 1] \neq \mathbb{E}[B_2(0,1) - B_2(0,0)|P = 0], \tag{14}$$

which leads to the result (7) that had to be shown.

C Re-evaluation of existing empirical evidence

This section contains the results of my meta-study of the papers studied in Maki et al. (2019). The results are presented in table 1 in the main text. Here, I motivate the conclusions for each of the evaluated papers.

- Carrico et al. (2018): condition 1 (asking participants to reduce consumption of red meet for environmental reasons) suggests that it is important to behave pro-environmentally. Thus, both desirability bias and crowding in of intrinsic motivation (and the associated direct effects of the policy on the nontargeted behavior) cannot be excluded.
- Geng et al. (2016): since there are only few observed individual characteristics, the CIA is not satisfied. The exclusion restriction appears plausible since it is hard to infer the intentions of the policy maker or the desirable behavior from a shopping list alone.
- Lacasse (2015): since there are only few observed individual characteristics, the CIA is not satisfied. In addition, questionnaire that aims at making salient past environmental behaviors provides direct cues for what the desirable answers/behaviors are (and has thus a direct impact on nontargeted behaviors/attitudes).

- Margetts and Kashima (2017): a survey with a rich set of questions is administered before the policy is implemented and before the pebs are performed. This allows using the CIA assumption. The policy itself can be interpreted as an ITT (for green shopping). Students put into the green supermarket may in principle infer from the large share of green products that the desired behavior is environmentally oriented. This would violate the exclusion restriction, so that a LATE would deliver biased results. In addition, as the authors implicitly acknowledge (when they discuss their study 2), the surveys in experiment 1a and 1b could be used by the participants to infer the purpose of the study, which potentially would induce a desirability bias.

- Lacasse (2019): a large number of pre-treatment covariates (demographic characteristics, preferences, past environmental behaviors) makes it potentially possible to use the study design and "match" on these characteristics.

- Lacasse (2016): there are only few pretreatment characteristics available, so that the CIA cannot be defended. In addition, asking about environmental behaviors and environmental identity potentially has a strong framing effect and an associated desirability bias (identical for both experiments in the paper)

- Parag et al. (2011): pro-environmental preferences are surveyed however after the targeted behavior, so that the CIA is violated (preferences are potentially an outcome of B_1).

- Poortinga et al. (2013): estimation results present evidence that is consistent with the policy having a direct effect on the intrinsic motivation/environmental identity. In particular, the authors measure the total effect of the (randomized policy) on surveyed environmental identity. As documented in the literature on crowing of intrinsic motivation (Bowles and Polania-Reyes, 2012), such an effect is consistent with a mechanism, in which the individual infers either about the desired behavior or about the intentions of the policy maker, which then leads to motivation crowding in. Although a survey measures individual characteristics, these cannot be used to identify the $\Delta^{spillover}$, because the timing prder of B_1 and B_2 cannot be established, and hence reverse causality cannot be excluded.

- Schultz et al. (2015): the study includes no observed covariates of individuals other than intentions to buy bulbs and their bulb-installing behaviors. Therefore, the CIA is not satisfied. The campaign targets energy saving via light bulbs and an exclusion restriction w.r.t. other environmental behaviors may be justified.

- Steinhorst et al. (2015): the pre-study includes a wider range of covariates, including attitudes. However, these were surveyed after the manipulation, which potentially creates endogeneity. In addition,

the pre-survey does not contain a randomized policy. In the main study, no observed covariates of individuals other than their targeted and nontargeted behavioral intentions were surveyed. Therefore, the CIA is not satisfied. The evidence presented in the paper is consistent with an experimental framing condition having a direct impact on the nontargeted intentions, e.g. through increasing the salience of the environmental aspects or through giving a clue of what the desired behavior is (desirability bias). The hypotheses are however stated as hypotheses regarding the direct effect of the policy $\Delta^{nontarget}$, even though the psychological explanations invoke a spillover effect $\Delta^{spillover}$. In particular, the paper invokes the explanation that the adopting the first behavior (as opposed to the policy) makes the norms salient, which then affects the second behavior.

- Thomas et al. (2016): this is the only study that actually estimates $\Delta^{spillover}$ and not $\Delta^{nontarget}$. However, the empirical design is not able to identify the timing of the targeted and nontargeted behaviors, so it is well possible that the nontargeted behaviors occur before the targeted behaviors (a reverse causality problem). In addition, there are not enough observed individual characteristics to control for endogenous selection (and hence, the CIA is not satisfied). As a result, the estimated coefficients represent estimates of the correlation between two behaviors Δ^{corr}. The exclusion restriction is potentially violated because of the environmental framing of the reform.

- Tiefenbeck et al. (2013): the study estimates the effect of the water report on electricity consumption (hence, $\Delta^{nontarget}$). The exclusion restriction is potentially violated because the policy suggests/argues that environmental behavior is desirable, and or because households might feel monitored. Thus, the negative effect is consistent with a direct effect of the policy on the nontargeted behavior as a result of crowding out of intrinsic motivation. The CIA cannot be established because of lack of observed individual characteristics.

- Truelove et al. (2016): The exclusion restriction could be deemed plausible because the action of the experimenter entails no clear hint (beyond the sign on the bin) that an environmental action is desirable. Had the noncompliers been kept in the study, a LATE approach would have yielded an estimate of $\Delta^{spillover}$. However, noncompliers are excluded from the study, so that the LATE approach cannot be applied on the remaining dataset. The CIA cannot be established because of lack of observed individual characteristics.

References

Alacevich, C., Bonev, P., and Söderberg, M. (2021). Pro-environmental interventions and behavioral spillovers: Evidence from organic waste sorting in sweden. *Journal of Environmental Economics and Management*, 108:102470.

Altmann, S., Grunewald, A., and Radbruch, J. (2022). Interventions and cognitive spillovers. *The Review of Economic Studies*, 89(5):2293–2328.

Bem, D. J. (1972). Self-perception theory. In *Advances in experimental social psychology*, volume 6, pages 1–62. Elsevier.

Blankenberg, A.-K. and Alhusen, H. (2018). On the determinants of pro-environmental behavior: A guide for further investigations. Technical report, CEGE Discussion Papers.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Brehm, S. S. and Brehm, J. W. (2013). *Psychological reactance: A theory of freedom and control*. Academic Press.

Brown, Z., Johnstone, N., Haščič, I., Vong, L., and Barascud, F. (2013). Testing the effect of defaults on the thermostat settings of oecd employees. *Energy Economics*, 39:128–134.

Bulte, E., List, J. A., and van Soest, D. (2021). Incentive spillovers in the workplace: Evidence from two field experiments. *Journal of Economic Behavior & Organization*, 184:137–149.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018). Putting your money where your mouth is: an experimental test of pro-environmental spillover from reducing meat consumption to monetary donations. *Environment and Behavior*, 50(7):723–748.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review*, 105(5):1–33.

Crudu, F., Knaus, M. C., Mellace, G., and Smits, J. (2022). On the role of the zero conditional mean assumption for causal inference in linear models. *arXiv preprint arXiv:2211.09502*.

Dolan, P. and Galizzi, M. M. (2015). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16.

Ek, C. (2018). Prosocial behavior and policy spillovers: A multi-activity approach. *Journal of Economic Behavior & Organization*, 149:356–371.

Ek, C. and Miliute-Plepiene, J. (2018). Behavioral spillovers from food-waste collection in swedish municipalities. *Journal of Environmental Economics and Management*, 89:168–186.

Festinger, L. (1957). A theory of cognitive dissonance (repr 1968).

Friedman, M. (1953). *Essays in positive economics*. University of Chicago press.

Galizzi, M. M. and Whitmarsh, L. (2019). How to measure behavioral spillovers: a methodological review and checklist. *Frontiers in psychology*, 10:342.

Geng, L., Cheng, X., Tang, Z., Zhou, K., and Ye, L. (2016). Can previous pro-environmental behaviours influence subsequent environmental behaviours? the licensing effect of pro-environmental behaviours. *Journal of Pacific Rim Psychology*, 10.

Goetz, A., Mayr, H., and Schubert, R. (2022). One thing leads to another: Evidence on the scope and persistence of behavioral spillovers. *Available at SSRN 3919454*.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Hansmann, R., Laurenti, R., Mehdi, T., and Binder, C. R. (2020). Determinants of pro-environmental behavior: A comparison of university students and staff from diverse faculties at a swiss university. *Journal of Cleaner Production*, 268:121864.

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*, 86:1.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874.

Hedlin, S. and Sunstein, C. R. (2016). Does active choosing promote green energy use: Experimental evidence. *Ecology LQ*, 43:107.

Hernán, M. A. (2016). Does water kill? a call for less casual causal inferences. *Annals of epidemiology*, 26(10):674–680.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (2010). Rubin causal model. In *Microeconometrics*, pages 229–241. Springer.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jessoe, K., Lade, G. E., Loge, F., and Spang, E. (2021). Spillovers from behavioral interventions: Experimental evidence from water and energy use. *Journal of the Association of Environmental and Resource Economists*, 8(2):315–346.

Jones, C. R., Whitmarsh, L., Byrka, K., Capstick, S., Carrico, A. R., Galizzi, M. M., Kaklamanou, D., and Uzzell, D. (2019). Methodological, theoretical and applied advances in behavioral spillover.

Korfiatis, K. J., Hovardas, T., and Pantis, J. D. (2004). Determinants of environmental behavior in societies in transition: evidence from five european countries. *Population and environment*, 25(6):563–584.

Lacasse, K. (2015). The importance of being green: the influence of green behaviors on americans' political attitudes toward climate change. *Environment and Behavior*, 47(7):754–781.

Lacasse, K. (2016). Don't be satisfied, identify! strengthening positive spillover by connecting pro-environmental behaviors to an "environmentalist" label. *Journal of Environmental Psychology*, 48:149–158.

Lacasse, K. (2019). Can't hurt, might help: Examining the spillover effects from purposefully adopting a new pro-environmental behavior. *Environment and Behavior*, 51(3):259–287.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, pages 43–58. Springer.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315.

Margetts, E. A. and Kashima, Y. (2017). Spillover between pro-environmental behaviours: The role of resources and perceived similarity. *Journal of Environmental Psychology*, 49:30 – 42.

McKenzie, C. R., Liersch, M. J., and Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, 17(5):414–420.

Nafziger, J. (2020). Spillover effects of nudges. *Economics Letters*, 190:109086.

Parag, Y., Capstick, S., and Poortinga, W. (2011). Policy attribute framing: A comparison between three policy instruments for personal emissions reduction. *Journal of Policy Analysis and Management*, 30(4):889–905.

Picard, J. (2022). Double-edged nudges? micro-foundations to behavioral interventions and their spillover effects. *Working paper*.

Poortinga, W., Whitmarsh, L., and Suffolk, C. (2013). The introduction of a single-use carrier bag charge in wales: Attitude change and behavioural spillover effects. *Journal of Environmental Psychology*, 36:240 – 247.

Raimi, K. T., Maki, A., Dana, D., and Vandenbergh, M. P. (2019). Framing of geoengineering affects support for climate change mitigation. *Environmental Communication*, 13(3):300–319.

Reichenbach, B. R. (1988). The law of karma and the principle of causation. *Philosophy East and West*, 38(4):399–410.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Schultz, P. W., Colehour, J., Vohr, J., Bonn, L., Bullock, A., and Sadler, A. (2015). Using social marketing to spur residential adoption of energy star®-certified led lighting. *Social Marketing Quarterly*, 21(2):61–78.

Singha, B., Karmaker, S. C., and Eljamal, O. (2023). Quantifying the direct and indirect effect of socio-psychological and behavioral factors on residential water conservation behavior and consumption in japan. *Resources, Conservation and Recycling*, 190:106816.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., and Greenberg, J. (2015). Understanding psychological reactance: new developments and findings. *Zeitschrift für Psychologie*, 223(4):205.

Steinhorst, J., Kloeckner, C. A., and Matthies, E. (2015). Saving electricity – for the money or the environment? risks of limiting pro-environmental spillover when using monetary framing. *Journal of Environmental Psychology*, 43:125 – 135.

Thøgersen, J. (1999). Spillover processes in the development of a sustainable consumption pattern. *Journal of economic psychology*, 20(1):53–81.

Thøgersen, J. (2004). A cognitive dissonance interpretation of consistencies and inconsistencies in environmentally responsible behavior. *Journal of environmental Psychology*, 24(1):93–103.

Thomas, G. O., Poortinga, W., and Sautkina, E. (2016). The welsh single-use carrier bag charge and behavioural spillover. *Journal of Environmental Psychology*, 47:126 – 135.

Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013). For better or for worse? empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy*, 57:160 – 171.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138.

Truelove, H. B., Yeung, K. L., Carrico, A. R., Gillis, A. J., and Raimi, K. T. (2016). From plastic bottle recycling to policy support: An experimental test of pro-environmental spillover. *Journal of Environmental Psychology*, 46:55 – 66.

VanderWeele, T. J. and Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of causal inference*, 1(1):1–20.