The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior

ETHAN BUENO DE MESQUITA   *University of Chicago*

SCOTT A. TYSON   *University of Rochester*

*W*e pose the commensurability problem: *When do the estimates generated by actual research designs correspond to quantities of theoretical interest? We study this question in settings where both treatment and outcome are behavior and the treatment effect of interest is decomposable into direct and informational channels. We establish two results. First, the quantity estimated by an actual research design is only commensurate with the total effect in the ideal experiment if treatment status in the research design is a sufficient statistic for the decision-makers' information. Second, a research design corresponding to a nonideal experiment isolates just the direct effect in the ideal experiment if two conditions hold: (i) there is no information effect in the nonideal experiment and (ii) the decision-maker's response function is additively separable in treatment and information. We apply our results to three substantive literatures: the efficacy of protest, the empowerment of female candidates, and indiscriminate violence in counterinsurgency.*

In many important social scientific settings, researchers are interested in estimating the effect of behavior on behavior. For instance, What is the effect of an increase in protests or violence by anti-government groups on government policy, be it concessions or repression (Collins and Margo 2007; Dell 2012; Gould and Klor 2010; Henderson and Brooks 2016; Hendrix and Salehyan 2012; Huet-Vaughn 2013; Madestam et al. 2013; Ritter and Conrad 2016)? What is the effect of a politician issuing press releases on voter behavior (Grimmer, Messing, and Westwood 2012)? What is the effect of one female candidate running for office on another female candidate's decision to run for office (Baskaran and Hessami 2018; Broockman 2014; Ladam, Harden, and Windett 2018)? Does indiscriminate violence by a counterinsurgent increase or decrease violence by rebels (Benmelech, Berrebi, and Klor 2014; Condra and Shapiro 2012; Dell and Querubín 2017; Jaeger et al. 2012; Lyall 2009)?

Although such questions are central, understanding what exactly we mean when we talk of the effect of one person's behavior on another person's behavior is something of a conceptual muddle, requiring careful analysis. This article is an attempt to contribute to that enterprise by articulating a framework and addressing some conceptual issues.

Despite the careful attention paid to identification issues by experimental methodologists, the conceptual difficulties for theoretical interpretability that arise in applications can be hard to spot precisely because it appears straightforward to describe such questions within the potential outcomes framework. We are trying to learn the effect of some agents' actions; call them the *treatment agents*. Call the agents whose actions are being affected the *outcome agents*. The set of treatments corresponds to the set of actions available to the treatment agents. And the potential outcomes correspond to the action each outcome agent would take under each possible action by the treatment agents. The causal effect of the treatment agents' actions on the actions of outcome agents is the difference in these potential outcomes under the different actions by the treatment agents.

With treatments and potential outcomes so defined, as emphasized by Angrist and Pischke (2009, chap. 1), we can then get clarity on the estimand of an empirical strategy by articulating an "ideal experiment." The ideal experiment again appears straightforward. One would like to randomly assign behavior by the treatment agents and observe the average response of the outcome agents under these different treatment assignments.

But the apparent simplicity of this formulation elides, rather than resolves, key conceptual issues. In many settings of interest, including all of those mentioned above, treatment behavior affects outcome behavior through at least two theoretical mechanisms: a *direct* channel and an *informational* channel. For instance, protest behavior might have a direct effect on government policy because larger protests are more disruptive, making governments more willing to take costly actions to bring them to an end. At the same time, protest behavior might also have an informational effect on government behavior because larger protests may change the government's beliefs about the amount of dissatisfaction among citizens, which might influence governmental policy choices. Similarly, indiscriminate bombing might have a direct effect on insurgency by killing or incapacitating rebels, while also conveying information about the counterinsurgents' resolve or level of concern for the welfare of the population.

The complication comes from the fact that the treatment is itself behavior, and people change their behavior

Ethan Bueno de Mesquita ⓘ , Harris School, University of Chicago, bdm@uchicago.edu.

Scott A. Tyson ⓘ , Department of Political Science, and W. Allen Wallis Institute of Political Economy, University of Rochester, styson2@ur.rochester.edu.

for a reason. Critically, the informational content of some change in the treatment agents' behavior may well depend on the reason behavior changed. So one must think carefully about how and whether an experimental manipulation maps onto some "real" reason.

At first blush, this conceptual problem appears less thorny when researchers achieve randomization of treatment behavior by exploiting randomization of nonbehavioral features of the world that influence behavior, rather than direct experimental manipulation. For instance, a researcher thinking about the effect of protest behavior on government policy might exploit experimental or quasi-experimental variation in protest behavior due to random shocks to the cost of protesting. But we still have to worry about whether different shocks generate different information, even if they induce the same change in treatment behavior. Suppose, for instance, that we are interested in the average treatment effect of an increase in protest size of 10,000 people. One possible source of increased protest size is a weather shock—e.g., it was sunny instead of rainy (Ritter and Conrad 2016). Another possible source is a shock to communication technology—e.g., some social media platform rolled out, allowing unfettered communication by potential protestors (Shapiro and Weidmann 2015). The government's inferences about the level of antigovernment sentiment may respond differently to two equally sized changes in protest behavior depending on which of these shocks is the source.

If different sources of variation in treatment behavior induce different information and, thus, different changes in outcome behavior, it may not make sense to talk about *the* effect of a change in the treatment agents' actions on the outcome agents' actions. Rather, we may have to talk about the effect of a change in the treatment agents' actions *due to some particular shock*.[1]

In light of this, articulating an appropriate ideal experiment requires a clear idea of exactly what one is trying to learn about. Suppose one wants to learn about the effect on government behavior of an increase in protest behavior that would result from a genuine increase in antigovernment sentiment. Then, the ideal experiment must surely involve a shock that influences protestors' (treatment agents) behavior, but which is not observable by the government (outcome agent). Such unobservable shocks come closest to representing the experimental ideal because, from the perspective of the government, the change in protest behavior might be the result of genuine changes to antigovernment sentiment. Of course, the research designs used to estimate causal effects in actual empirical work sometimes do not exploit this type of unobservable shock. Again, for instance, in the literature on the efficacy of protest, a common research design involves using shocks to weather as a source of random variation in protest behavior. But the government presumably observes the weather.

This last observation points to a novel set of questions, which we collectively refer to as the *commensurability problem*.[2] Do standard research designs used in various literatures assessing the effect of behavior on behavior in fact generate estimates that are interpretable in terms of some quantity of theoretical interest in an ideal experiment? That is, under what conditions are actual research designs and ideal experiments theoretically commensurate? And, if they are, what theoretical quantity in the ideal experiment is the research design estimating?

To make progress on those questions, we propose a theoretical framework in which one set of agents' behavior might affect another set of agents' behavior through both direct and informational channels. In that framework, following the argument above, we think of the ideal experiment as involving shocks to some feature of the world that influences the treatment agents' behavior, but which is not observable by the outcome agents. We define the average total effect of a change in treatment behavior on outcome behavior in that ideal experiment. Moreover, we decompose this total effect in the ideal experiment into a direct effect and an informational effect.

We can also study nonideal experiments in our theoretical framework to represent a variety of actual research designs. The difference in outcome behavior under different (shock-induced) treatment behaviors in the nonideal experiment represents what a researcher actually observes in an empirical study using a research design that corresponds to the nonideal experiment.

With this formalism in hand, we ask two questions. First, under what conditions is a nonideal experiment commensurate with the ideal experiment for the total average treatment effect (i.e., the combined direct and informational effects)? By this, we mean to imagine shocks that induce the same magnitude of change in treatment behavior in the ideal experiment and some nonideal experiment. When will the total effect in the nonideal experiment (representing the estimates generated by some actual research design) be the same as the total effect in the ideal experiment? Unfortunately, as we show in our first theorem, the answer is never.[3] Thus, with respect to the total effect, there is a fundamental commensurability problem. So, for instance, the effect of changes in protest behavior estimated from weather shocks is an inherently different quantity than the effect of changes in protest behavior due to, say, unobservable changes in citizens' antigovernment sentiment or even unobservable shocks to the cost of protesting.

The example of weather shocks provides one way of motivating the intuition for our second question. When

---

[1] This is not a point about heterogeneous treatment effects. In both cases, we are changing the treatment (e.g., the amount of protest) by the exact same amount and studying the response of a single decision-maker.

[2] In philosophy of science, commensurability problems typically refer to situations in which two theories are conceptually incompatible—that is, the concepts operating in one theory cannot be coherently translated into the concepts of another theory (Feyerabend 1962; Kuhn 1962). In an homage to that idea, we use the term to refer to a setting in which effects as defined in a theory and effects as defined in terms of potential outcomes for the purposes of an empirical exercise are incompatible.

[3] More precisely, the effect estimated by the nonideal experiment is the same as the total effect in the ideal experiment on at most a set of measure zero.

governments observe a large protest on a sunny day, they know that it conveys different information than a large protest on a rainy day. Indeed, maybe there is no information effect when shocks are observable. Perhaps, the outcome agent, having observed the shock to treatment behavior, doesn't update her beliefs at all ("I know those extra 10,000 people are here just for the sun"). In this case, perhaps the effect identified by a nonideal experiment is commensurate with the direct effect from the ideal experiment.

Exploring this possibility leads to our second theorem, where the news is more positive. A nonideal experiment isolates just the direct effect in the ideal experiment under two conditions:

1. There is no informational component in the effect from the nonideal experiment.
2. The direct and informational mechanisms in the ideal experiment are additively separable in the outcome agents' response.

In what follows, we start with a simple example of a specific model of protest efficacy to illustrate the basic intuitions of our results. We then show that these intuitions hold in a general framework and, consequently, that commensurability problems apply across a wide range of substantive settings.

To illustrate our results' applicability, we use them to discuss three substantive literatures in economics and political science. Our first application connects with our motivating example involving the literature assessing the effects of protests or violence by citizens on government policy (Collins and Margo 2007; Dell 2012; Henderson and Brooks 2016; Hendrix and Salehyan 2012; Huet-Vaughn 2013; Madestam et al. 2013). Here, we show that there is a fundamental commensurability problem—even work using research designs that make real progress on causal identification do not yield estimates that are interpretable as quantities of theoretical interest in the ideal experiment.

Our second application corresponds to empirical work on the effect of female electoral victory on the empowerment of future female candidates (Baskaran and Hessami 2018; Broockman 2014; Ladam, Harden, and Windett 2018). Here, we show that recent work using the election regression discontinuity design is again not capturing the total effect of female electoral victory from the ideal experiment. However, our results suggest the possibility that this work is isolating the direct effect and highlights the key substantive question whose answer determines whether or not this is the case.

Our third application corresponds to work on the effects of indiscriminate violence by counterinsurgents on rebel violence (Benmelech, Berrebi, and Klor 2014; Condra and Shapiro 2012; Dell and Querubín 2017; Jaeger et al. 2012; Lyall 2009). We focus on the recent work by Dell and Querubín (2017) which, our results suggest, use a research design that may in fact estimate the total effect in the ideal experiment.

Overall, we believe that the framework we develop, the results we derive, and the applications we explore suggest that grappling with the commensurability problem is indeed important for the theoretical interpretation of even well-identified empirical results in settings where treatment results from people's behavior and behavior potentially conveys information. It is our hope that, by emphasizing the commensurability problem and offering some conceptual clarification of the issues it raises, our analysis will help with ongoing efforts to bridge the gap between theoretical models and empirical scholarship in the causal inference tradition.

RELATIONSHIP TO EXISTING LITERATURE

In important articles, Deaton (2010) and Heckman and Urzua (2010) argue that the choice of an instrument necessarily restricts an empirical research question. They focus on when the local average treatment effect (LATE) is interpretable in terms of the structural parameters of a theoretical model. These two articles essentially present an informal version of what we might describe as a commensurability problem between the LATE obtained in a standard reduced-form approach and structural parameters.[4] Deaton (2010) and Heckman and Urzua (2010) argue that "incommensurability" between the LATE and a structural parameter arises because of cross-sectional heterogeneity in treatment effects, specifically, when the response to treatment is not the same across subpopulations.

Our results differ from this point in two important ways. First, since we focus on a setting with a single decision-maker, there is no cross-sectional heterogeneity in treatment effects in our model. Hence, the source of the commensurability problem we identify is distinct from those discussed in the earlier literature. In particular, our commensurability problem has nothing to do with localness—it has to do with the way in which the same variation in treatment results in different informational effects depending on the source of that variation.

Second, we are not focused on whether a reduced-form approach can identify structural parameters. Deaton (2010) and Heckman and Urzua (2010) criticize reduced-form approaches by arguing for a different estimand, and in this sense, they are rejecting the goals of identification-oriented empirical researchers. By contrast, we concede that the estimands associated with reduced-form relationships can be of theoretical interest. We then show conditions under which standard research designs employed in various substantive literatures do or do not capture these very estimands from the ideal experiment.

The work most similar to ours represents a specific research design in a formal model and asks what quantity it recovers. For instance, Eggers (2017) directly represents the election regression discontinuity (RD) in a model of the incumbency advantage to show that the RD does not purge electoral selection. Fudenberg and

---

[4] See Goldberger (1972) for a description of structural models, and Angrist, Imbens, and Rubin (1996) for a discussion of the relationship between structural models and the potential outcome framework.

Levine (2019) model research designs, such as RD or difference-in-differences, in a model where information feedback affects effort that endogenously determines treatment. Other work, somewhat further afield, uses theory to think about how to optimally design treatments to maximize learning (Banerjee, Chassang, and Snowberg 2017; Chassang, Padró i Miquel and Snowberg 2012) or to question the normative or positive interpretation of well-identified empirical findings (Ashworth and Bueno de Mesquita 2014; Ashworth, Bueno de Mesquita, and Friedenberg 2018; Fowler 2018; Izzo, Dewan, and Wolton 2018; Prato and Wolton 2018; Sun and Tyson 2019; Wolton 2019).

A SIMPLE ILLUSTRATIVE EXAMPLE

To build intuition, we start with a simple formal example that illustrates our two key results, which are then shown in greater generality in the sequel. This example, unlike our theorems, is focused on the protest application and is built on specific functional forms.

A group of citizens decides whether to protest, and then, the government responds. To keep things simple, we treat the group of citizens as a unitary actor. The general framework in the next section however does not require this sort of simplification.

The group has one of two types $\theta \in \{\underline{\theta}, \bar{\theta}\}$. Think of the group's type as representing its level of antigovernment sentiment. The prior probability that $\theta = \bar{\theta}$ is $p \in (0, 1)$. The cost of protesting may be low ($\underline{c}$) or high ($\bar{c}$), each with equal probability. High costs can be thought of as representing, for example, bad weather.

The group protests if and only if $\theta > c$. Assume $\bar{\theta} > \bar{c} > \underline{\theta} > \underline{c}$, so that a high type always protests and a low type protests if and only if the costs of protesting are low. So more antigovernment types are willing to bear higher costs to protest than are less antigovernment types.

The government observes protest behavior and forms a posterior belief, $\hat{p}$, about the probability that $\theta = \bar{\theta}$. The government's response is described by a function r that takes as an input whether there was a protest and the government's posterior belief. Let $\mathbb{I}$ be an indicator function that takes the value 1 if there was a protest and 0 if there was not. Then, the government's response function is as follows:

$$r(\mathbb{I}, \hat{p}) = \begin{cases} \bar{\alpha} + \bar{\beta} \cdot \hat{p}^\gamma & \text{if } \mathbb{I} = 1 \\ \underline{\alpha} + \underline{\beta} \cdot \hat{p}^\gamma & \text{if } \mathbb{I} = 0, \end{cases}$$

with $\bar{\alpha} > \underline{\alpha} > 0$, $\bar{\beta} \geq \underline{\beta} > 0$, and $\gamma > 0$.

The parameter α represents the government's direct response to having protestors in the street—for instance, the government may be willing to make concessions or engage in repression to end a disruptive protest. So $\bar{\alpha} > \underline{\alpha}$ says that protests directly influence government behavior. The term $\beta \cdot \hat{p}^\gamma$ represents the government's response to its beliefs about the citizens' level of antigovernment sentiment—for instance, the government's willingness to make concessions or engage in repression may depend on how disgruntled it believes its citizen are. If $\bar{\beta} > \underline{\beta}$, there is an interaction between the direct and informational inputs

to the government's response—for instance, the government may be more willing to engage in repression to end a disruptive protest when it is particularly worried about citizen disgruntlement. By contrast, if $\bar{\beta} = \underline{\beta}$, then the direct and informational effects of protest on government behavior are additively separable. The term $\hat{p}^\gamma$ says that the government's response is increasing in its posterior beliefs that $\theta = \bar{\theta}$. The parameter γ allows flexibility in the shape of this function.

Representing the Ideal Experiment: Unobservable Costs

An ideal experiment would manipulate protest behavior by changing the costs in a way that is unobservable to the government. So imagine a situation in which c cannot be observed by the government but can be observed by the citizen group and a researcher.

Using Bayes' rule, if a protest occurs, the government has posterior beliefs:

$$\Pr(\bar{\theta} \mid \text{protest}) = \frac{2p}{1+p}.$$

If a protest does not occur, the government is certain the group is of low type:

$$\Pr(\bar{\theta} \mid \text{no protest}) = 0.$$

What is the average effect on the government's response of a decrease in the cost of protesting?

If costs are low, then a protest occurs for sure and the government's response is $r\left(1, \frac{2p}{1+p}\right)$. If costs are high, then a protest occurs with probability p. If a protest occurs, the government's response is again $r\left(1, \frac{2p}{1+p}\right)$. But if a protest does not occur, the government's response is $r(0, 0)$. So the average effect of lower costs is

$$r\left(1, \frac{2p}{1+p}\right) - \left[pr\left(1, \frac{2p}{1+p}\right) + (1-p)r(0,0)\right]$$
$$= (1-p)\left[\bar{\alpha} - \underline{\alpha} + \bar{\beta} \cdot \left(\frac{2p}{1+p}\right)^\gamma\right]. \quad \textbf{(1)}$$

Equation (1) represents the reduced form causal relationship between the costs of protesting and the government's behavior (think of the analogue to two-stage least squares). Of course, lowering the cost does not always result in a change in protest behavior. With low costs, there is always protest, whereas with high costs, there is protest with probability p. Consequently, the effect of lowering of costs on the probability of protest is $1 - p$. Dividing the reduced-form relationship in Equation (1) by this "first stage" gives the local total average effect of increased protest on government response that results from unobservably lowering costs (this is the analogue of the IV or Wald estimator). We label this τ_u:

$$\tau_u = \bar{\alpha} - \underline{\alpha} + \bar{\beta} \cdot \left(\frac{2p}{1+p}\right)^\gamma. \quad \textbf{(2)}$$

We can decompose this total effect into two substantive components: a direct effect and an informational effect. The direct effect is the effect of

increased protest behavior holding beliefs fixed (here we hold them fixed at $\frac{2p}{1+p}$):

$$\delta_u = \bar{\alpha} - \underline{\alpha} + \left(\bar{\beta} - \underline{\beta}\right) \cdot \left(\frac{2p}{1+p}\right)^\gamma. \qquad (3)$$

The informational effect is the effect of changing the government's beliefs from what they would be if no protest occurs (0) to what they would be if a protest occurs $\left(\frac{2p}{1+p}\right)$, holding protest behavior fixed (here, fixed at a protest not occurring):

$$\iota_u = \underline{\beta}\left(\frac{2p}{1+p}\right)^\gamma. \qquad (4)$$

The total effect (τ_u) is equal to the sum of the direct effect (δ_u) and informational effect (ι_u). Notice, we could have defined the direct effect holding beliefs fixed instead at 0 and then redefined the information effect at a protest occurring. Nothing hinges on this choice.

Importantly, $\tau_u, \delta_u,$ and ι_u do not represent quantities that we think of as directly observable by an empirical researcher. Rather, these are representations of effects of theoretical interest *in the ideal experiment*, which an empirical researcher might never encounter. The commensurability problem is about whether actual research designs estimate any of these quantities.

Representing an Actual Research Design: Observable Costs

A common research design uses weather as an instrument, which of course, is observable to the government. So our next step is to study, within our model, the effect of an observable shock to the cost of protesting. Then, we ask whether there is a commensurability problem: that is, does the quantity that the actual research design estimates correspond to a theoretical quantity of interest in the ideal experiment?

So let's think about the same model, but now assume that the costs are observable to the government. Using Bayes' rule again, we now have to calculate the government's beliefs conditional on both whether or not a protest occurs and whether costs are high or low. If costs are low, both types protest, so the government learns nothing and its posterior equals its prior:

$$\Pr(\bar{\theta} \mid \text{protest}, \underline{c}) = p.$$

By contrast, if costs are high, one type protests and the other doesn't, so the government learns everything:

$$\Pr(\bar{\theta} \mid \text{protest}, \bar{c}) = 1.$$

$$\Pr(\bar{\theta} \mid \text{no protest}, \bar{c}) = 0.$$

What is the average effect on the government's response of the increased probability of protest associated with a decrease in the cost of protesting when costs are observable?

If costs are low, then a protest occurs for sure and the government's response is $r(1, p)$. If costs are high, then a protest occurs with probability p. If a protest occurs, the government's response is $r(1, 1)$, but if a protest does not occur, the government's response is $r(0, 0)$. So the average effect of lowering observable costs is

$$r(1, p) - [pr(1, 1) + (1 - p)r(0, 0)]$$
$$= (1 - p)(\bar{\alpha} - \underline{\alpha}) + \bar{\beta}(p^\gamma - p).$$

Just as before, we divide by $1 - p$ to get the effect of the change in protest behavior due to a change in costs (τ_o):

$$\tau_o = \bar{\alpha} - \underline{\alpha} + \bar{\beta} \cdot \frac{p^\gamma - p}{1 - p}.$$

If an empirical researcher uses a research design that corresponds to this nonideal experiment (i.e., one with observable cost shocks), then τ_o represents the resulting IV estimate.

Commensurability

It is now straightforward to think about commensurability in the context of our example. We imagine that an empirical researcher uses a research design with observable costs. Thus, the only thing she actually observes is an estimate of τ_o. But what she really wants to learn about are effects of theoretical interest in the ideal experiment—τ_u, δ_u, or ι_u. Can she use her estimate of τ_o to do so?

A first observation is that the total effect of the change in protest induced by a shock to costs is not the same when that shock is and is not observable. Formally, $\tau_u = \tau_o$ if and only if

$$\left(\frac{2p}{1+p}\right)^\gamma = \frac{p^\gamma - p}{1 - p},$$

which never holds. The left-hand side is greater than the right-hand side for any $p \in (0, 1)$ and $\gamma > 0$. This means that the research design using observable weather shocks does not estimate the total effect in the ideal experiment, a manifestation of the commensurability problem in our example.

The key thing going on is a difference in the information effects between the actual research design and the ideal experiment. In both cases, there is the same change in protest behavior in response to a cost shock—the probability of protest increases by $1 - p$. But there are different effects on beliefs. When costs are observable, the informational content of a protest depends on what the costs are—the government understands that a large protest means something different on a sunny day and on a rainy day. When costs are low, there is no informational content—holding a protest on a nice sunny day is too easy for the government to conclude anything about citizen attitudes. But when costs are high, there is considerable informational content because only truly angry citizens protest in the rain. By contrast, with unobservable shocks, the government has to average across these possibilities, leading to different beliefs in the two scenarios. As a consequence, the same change in average protest behavior has different effects on government behavior in the ideal experiment and the actual research design. This creates the commensurability problem: the quantity estimated by the research design does not correspond to the theoretical object of interest in the ideal experiment. This fact corresponds to our first theorem below.

One way of thinking about this result is that an observable shock violates the exclusion restriction because it affects the way the treatment (here, protest) is interpreted. In this sense, our argument can be understood as illustrating how theoretical arguments can highlight an important way in which, to use Heckman's (2000) terminology, an external instrument may fail to be exogenous. (Sarsons [2015] discusses other ways rainfall may violate the exclusion restriction.)

A thought one might have is that, although the estimate generated by a research design with observable shocks does not correspond to the *total* effect of increased protest in the ideal experiment, maybe it corresponds to just the direct effect. The intuition is that when the government observes the cost shock, it knows those shocks are causing a change in protest behavior. Hence, it won't mistakenly think that increased protest size due to lower costs is the result of a genuine change in antigovernment sentiment. So, maybe once it takes the shock into account, there is no informational content left in the change in protest behavior. If this is the case, perhaps the effect being estimated by a research design using observable shocks to protest costs captures just the direct effect in the ideal experiment, purged of any informational effect.

Comparing the total effect in the model with observable shocks to just the direct effect in the model with unobservable shocks, and rearranging, we see that they are equal if and only if

$$\frac{(1+p)^\gamma (p^\gamma - p)}{(2p)^\gamma (1-p)} = \frac{\bar{\beta} - \underline{\beta}}{\bar{\beta}}.$$

The left-hand side of this condition is strictly decreasing in γ. The right-hand side is constant in γ. As such, for any $p \in (0, 1)$, this condition holds for at most one value of γ, which is to say that the actual research design isolates the direct effect in the ideal experiment only in a knife-edge case.

To start to get some intuition for what is going on, think about the case where the direct and informational effects of protest on government behavior are additively separable (i.e., $\bar{\beta} = \underline{\beta}$). In this case, the right-hand side is zero and the unique γ for which the condition holds is $\gamma = 1$. That is, if the direct and informational effects of protest on government behavior are additively separable, then the total effect in the actual research design equals the direct effect in the ideal experiment if and only if the government's behavior is linear in its beliefs. Why is this true?

Additive separability implies that the direct effect doesn't depend on the beliefs. This means that the direct effect in the actual research design and the ideal experiment are the same, even though the government always has different beliefs in those two settings. As such, for the total effect in the actual research design to be equal to the direct effect in the ideal experiment, there must be no information effect in the actual research design. It might seem that this is impossible because the government does learn from observing protest size. But, as must be the case for a Bayesian actor, the government's posterior beliefs *on average* equal its prior beliefs. And so the information effect, which averages across the possible shocks, is zero if and only if beliefs enter linearly, so that the average is all that matters.

If there is not additive separability (i.e., $\bar{\beta} \neq \underline{\beta}$), then there is in fact heterogeneity in the direct effect itself because it depends on the government's beliefs.[5] So now the direct effects in the actual research design and the ideal experiment are not equal. As such, for the total effect in the actual research design to be equal to the direct effect in the ideal experiment, it must be that the information effect in the actual research design is of precisely the right size and sign to exactly off-set the difference between the direct effects in the actual research design and in the ideal experiment. This is obviously a knife-edge condition, reflected in the fact that it holds for one and only one value of γ.

Our second theorem establishes that this logic generalizes. The quantity estimated by some research design in which the government's information differs in any way from the ideal experiment only corresponds to the direct effect in the ideal experiment if: (1) there is no informational effect in the actual experiment and (2) the government's best response is additively separable in the direct effect of protest and it's beliefs about the state of the world.

GENERAL FRAMEWORK

The example illustrates the key forces at work in our results. However, it was special, making use of specific functional forms and other restrictive assumptions. Moreover, it was directly tied to the protest interpretation. In this section, we develop a more general and flexible framework to explore the commensurability problems highlighted by the example. Proving similar results in this setting will elucidate the ways in which the commensurability problem does not depend on these special assumptions and thus applies across a wide array of substantive applications.

In our formalization, there are three random variables: a state of the world (θ), a shock (ω), and idiosyncratic noise (ε). The state of the world θ is drawn from the set $\{\underline{\theta}, \bar{\theta}\}$ according to the prior distribution $p = \Pr(\bar{\theta})$. The shock is drawn from a compact set $\Omega \subset \mathbb{R}$. Let $(\Omega, \mathcal{B}, \mu)$ be a probability space with Borel σ-algebra, $\mathcal{B}$, and probability measure μ that is absolutely continuous with respect to Lebesgue measure. The idiosyncratic noise is a random variable that is independent of everything. This introduces observation noise into the process and implies that the decision-maker's posterior beliefs are never degenerate. The only difference between the shocks and observation noise is that the empirical researcher also does not observe this idiosyncratic noise; our results do not rely on the presence of this idiosyncratic noise.

---

[5] This heterogeneity is distinct from the cross-sectional heterogeneity highlighted by Deaton (2010).

These three factors, together, determine the outcome of a "social process" (e.g., a protest or election) according to the following:

$$A_{\theta,\varepsilon}^{\omega} = F(\theta + \omega + \varepsilon),$$

where $F : \mathbb{R} \to \mathbb{R}$ is a strictly monotone function. The social process could be the result of strategic considerations of a set of individuals, along with the associated equilibrium logic, but our results do not require on such a formulation. The social process determines *treatment assignment* according to

$$T_{\theta,\varepsilon}^{\omega} \equiv T\left(A_{\theta,\varepsilon}^{\omega}\right).$$

The function T is weakly monotone and maps from outcomes of the social process to a linearly ordered set of possible treatments, $\mathcal{T}$.

Let $I(\theta, \omega, \varepsilon)$ be the decision-maker's information set. Then, her posterior belief is

$$\pi(I(\theta, \omega, \varepsilon)) = \Pr(\theta = \bar{\theta} \,|\, I(\theta, \omega, \varepsilon)).$$

Notice, for example, that if the shock is unobservable by the decision-maker, then this posterior belief might depend on ω only through its effect on A or T. However, if ω is observable, then the posterior belief might be different for two different values of the shock, ω, even at fixed values of A and T. This observation will become important below.

The outcome of interest for the empirical analyst is the decision-maker's *response*. The set of possible responses is a compact subset $\mathcal{R} \subset \mathbb{R}$. The decision-maker's response depends on the value of treatment as well as the decision-maker's posterior beliefs about the state of the world. It is represented by the one-to-one and continuously differentiable function:

$$r(T, \pi) : \mathcal{T} \times [0, 1] \to \mathcal{R}.$$

Microfounding the decision-maker's response as coming from the maximization of a utility function is standard and so we omit the details (see, e.g., Mas-Collel, Whiston, and Green [1995, chap. 3] or Kreps [2012, chap. 3]).[6]

Contrasts and Effects

The shock plays a critical role in our analysis, specifically, its role is to exogenously vary the outcome of the social process, A, thus changing treatment independently of the state of the world θ. We start by using the shock to define causal effects, but before doing so, it will be useful to introduce a little terminology.

Definition 1. *A **contrast** is a pair (ω', ω'') where $\omega', \omega'' \in \Omega$ such that $\omega' \neq \omega''$. The set of contrasts is*

$$\mathcal{C} = \{(\omega', \omega'') \in \Omega \times \Omega \,|\, \omega' \neq \omega''\},$$

endowed with the product measure.

We assume that shocks matter for treatment on average. That is, there must be a "first stage relationship" or there is no effect to study. So, for all (ω', ω''),

$$\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right] \neq 0.$$

A contrast comprises two distinct values of the shock. The local average total effect of a change in treatment due to a contrast is obtained by comparing the decision-maker's average response at two distinct values of the shock divided by the average value of treatment at the same two values of the shock:

Definition 2. *The **local average total effect** at the contrast $(\omega', \omega'') \in \mathcal{C}$ is*

$$\tau(\omega', \omega'')$$
$$= \frac{\mathbb{E}_{\theta,\varepsilon}\left[r\left(T_{\theta,\varepsilon}^{\omega''}, \, \pi(I(\theta, \omega'', \varepsilon))\right) - r\left(T_{\theta,\varepsilon}^{\omega'}, \, \pi(I(\theta, \omega', \varepsilon))\right)\right]}{\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right]}.$$

The total effect can be decomposed into a direct effect and an informational effect.

Definition 3. *Fix a posterior belief $\pi_{\theta,\varepsilon}^{\omega'} = \pi(I(\theta, \omega', \varepsilon))$. The **local average direct effect** at the contrast (ω', ω'') and the belief $\pi_{\theta,\varepsilon}^{\omega'}$ is*

$$\delta(\omega', \omega'') = \frac{\mathbb{E}_{\theta,\varepsilon}\left[r\left(T_{\theta,\varepsilon}^{\omega''}, \pi_{\theta,\varepsilon}^{\omega'}\right) - r\left(T_{\theta,\varepsilon}^{\omega'}, \pi_{\theta,\varepsilon}^{\omega'}\right)\right]}{\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right]}.$$

Definition 4. *Fix a treatment status $T_{\theta,\varepsilon}^{\omega''}$. The **local average information effect** at the contrast (ω', ω'') and the treatment status $T_{\theta,\varepsilon}^{\omega''}$ is*

$$\iota(\omega', \omega'')$$
$$= \frac{\mathbb{E}_{\theta,\varepsilon}\left[r\left(T_{\theta,\varepsilon}^{\omega''}, \pi(I(\theta, \omega'', \varepsilon))\right) - r\left(T_{\theta,\varepsilon}^{\omega''}, \pi(I(\theta, \omega', \varepsilon))\right)\right]}{\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right]}.$$

From Definitions 2–4, we have the following decomposition:

$$\tau(\omega', \omega'') = \delta(\omega', \omega'') + \iota(\omega', \omega'') \qquad (5)$$

$$= -(\delta(\omega'', \omega') + \iota(\omega'', \omega')). \qquad (6)$$

Notice that for a single contrast, (ω', ω''), there are two (potentially) different direct and informational effect pairs whose sum corresponds to the same total effect. Nothing hinges on which pair one chooses.

Experiments

Our definition of an experiment is in the spirit of Blackwell (1951). Our framework describes a set of states corresponding to the cross-product of the sets of contrasts, states of the world, and idiosyncratic noise. In our terminology, an *experiment*, $\mathcal{E}$, is the probability measure over this set of states, along with the decision-maker's information set (which we denote $I^{\mathcal{E}}$). When we compare experiments (representing various research designs), we hold the probability measure over the set of states fixed and vary the decision-maker's information set.

---

[6] See Conlisk (1973) on functional forms of response functions in experimental studies.

To start thinking about commensurability, we first need to define the *ideal experiment.* As we discussed in the introduction, and illustrated in our example, an ideal experiment is one where the decision-maker observes treatment but does not observe the process that leads to her particular treatment assignment.

Definition 5. *The **ideal experiment** is an experiment where the decision-maker's information set is* $I^{\mathcal{I}}(\theta, \omega, \varepsilon) = \left\{ T^{\omega}_{\theta, \varepsilon} \right\}$.

Commensurability

We are interested in whether various nonideal experiments, corresponding to canonical research designs used in prominent empirical literatures, might estimate an effect of theoretical interest in the ideal experiment. To answer this question, we first need to develop some terminology.

Label the total, direct, and information effects at a contrast (ω', ω'') in some experiment $\mathcal{E}$ by $\tau^{\mathcal{E}}_{\theta, \varepsilon}(\omega', \omega'')$, $\delta^{\mathcal{E}}_{\theta, \varepsilon}(\omega', \omega'')$, and $\iota^{\mathcal{E}}_{\theta, \varepsilon}(\omega', \omega'')$, respectively. For the ideal experiment in particular, we label these quantities $\tau^{\mathcal{I}}_{\theta, \varepsilon}(\omega', \omega'')$, $\delta^{\mathcal{I}}_{\theta, \varepsilon}(\omega', \omega'')$, and $\iota^{\mathcal{I}}_{\theta, \varepsilon}(\omega', \omega'')$.

Conceptually, one can think of any given study as being represented by one realization of some experiment in our framework. So we start by defining what it means for one particular instance of an experiment, associated with some particular contrast, to capture an effect of interest in the ideal experiment. The idea is that the experiment captures the effect in the ideal experiment if the total effect in the experiment (which represents the only thing the empirical researcher observes) equals the effect of interest in the ideal experiment.

Definition 6. *An experiment,* $\mathcal{E}$, *captures an effect* $E^{\mathcal{I}} \in \left\{ \tau^{\mathcal{I}}, \delta^{\mathcal{I}}, \iota^{\mathcal{I}} \right\}$ *in the ideal experiment at the contrast* (ω', ω'') *and prior p if*

$$\tau^{\mathcal{E}}(\omega', \omega'') = E^{\mathcal{I}}(\omega', \omega'').$$

Of course, an experiment capturing an effect of theoretical interest in the ideal experiment at one particular contrast and prior is not enough to have confidence that we are learning about the quantity of interest. For that, we want a theoretical analogue of econometric identification (Imbens and Angrist 1994; Koopmans and Reiersol 1950). That is, we want to ensure that an actual research design captures the effect of interest in the ideal experiment with probability one.

Definition 7. *An experiment,* $\mathcal{E}$, *is **commensurate** with the ideal experiment for an effect* $E^{\mathcal{I}} \in \left\{ \tau^{\mathcal{I}}, \delta^{\mathcal{I}}, \iota^{\mathcal{I}} \right\}$ *if* $\mathcal{E}$ *captures the effect* $E^{\mathcal{I}}$ *in the ideal experiment at almost every contrast and prior* $\langle (\omega', \omega''), p \rangle \in \mathcal{C} \times (0, 1)$.

Discussion of Concepts

In this section, we discuss two important components of our analysis. First, we discuss our choice to study the commensurability problem in the context of an ideal experiment. Second, we briefly relate our assumptions to common theoretical frameworks.

Haavelmo (1944, 14) distinguishes between two different classes of experiments, an ideal hypothetical experiment that isolates causal channels and the set of experiments that are naturally produced by Nature. Our analysis is about the relationship between these two types of experiments. Building on this classic distinction, the benchmark of the ideal experiment is endorsed by Angrist and Pischke (2009, 1) as a way of precisely articulating causal questions, with a particular emphasis on assessing the impact of interventions.

An important challenge to causal studies is articulated by Deaton (2010) and Heckman (2000, 84–6), who stress the importance of using a structural model as a benchmark to evaluate the local average treatment effect because, they argue, structural parameters have a clear substantive interpretation. They further argue that because most studies do not link the local average treatment effect to a structural model, a theoretical interpretation of the LATE is not implied.

We, of course, agree that learning about structural parameters is the most satisfying version of estimating quantities of theoretical interest. Our approach, however, allows for the possibility that there are reduced-form quantities, short of structural parameters, that might nonetheless qualify as of theoretical interest. For instance, in a study of protests, one might want to answer a question like, "if the leader of an antigovernment group were to switch to using violent tactics, would the movement become more or less effective?" Obviously, if one knew the structural parameters of the right model of protest, one could answer this question. But this is more demanding than is needed. One could also answer this question with the total effect from an ideal experiment. Hence, we define the ideal experiment and the associated effects as we do to give the reduced-form approach the best shot possible at identifying quantities of theoretical interest even when it cannot identify structural parameters.

To highlight our results, we develop our framework abstractly, rather than providing a more detailed behavioral or game-theoretic structure. We do this for a number of reasons. First, our results follow from general properties, namely, that the decision-maker responds to treatment status and beliefs.[7] Second, our results are consistent with a number of different substantive scenarios. Generality, emphasize that many different game forms are consistent with our approach. Finally, deriving these features of the environment as part of a game form would involve a number of additions to our setup, such as payoffs and solution concepts, all of which are standard and would not influence our main results, but would complicate exposition.

MAPPING THE FRAMEWORK TO APPLICATIONS

Before turning to the main analysis, we show how three important substantive applications, which have each

---

[7] For instance, we do not require the decision-maker to update her beliefs consistently with Bayes' rule, but simply that her beliefs are not constant across different pieces of information.

generated attention by empirical researchers interested in causal effects, fit into our framework.

The Efficacy of Protest

Protest and other forms of mobilization can influence government policy—be it repression, concessions, or other responses—through a multitude of different channels, many of which fit into the distinction between the direct and informational channels highlighted above. For instance, mobilized dissent, as an expression of political preferences, can communicate information to the government about the extent of political dissatisfaction in society, and, thus, may have an informational effect on the government's response. In addition, protests may be disruptive to the functioning of society or the economy, and consequently, governments may respond to such dissent even in the absence of new information about citizen dissatisfaction, thus giving rise to a direct effect.

Assessing the efficacy of various forms of political mobilization is a central, and vexing, concern for empirical studies. To assess the efficacy of political mobilization, one needs to compare government responses with various levels or types of mobilization. The fundamental problem, of course, is that the decision to mobilize is endogenous.

Not surprisingly, empirical work has turned to searching for sources of exogenous variation in political mobilization. The most prevalent strategy is an instrumental variables approach (Collins and Margo 2007; Dell 2012; Henderson and Brooks 2016; Hendrix and Salehyan 2012; Huet-Vaughn 2013; Madestam et al. 2013; Ritter and Conrad 2016). The basic idea is to look for factors that shift individuals' propensity to mobilize, but that have no independent channel to affect outcomes. The most common source of variation used in this literature comes from the weather. The idea is that if a protest is scheduled for a certain date, and on that date there is, say, unexpected torrential rain or unusually high temperatures, the individual cost of participating in the protest increases. Thus, these random natural phenomena provide an exogenous shock to citizens' costs of participating in a protest, which can be used in an instrumental variables analysis to isolate the effect of mobilized turnout on government behavior.

Representing this approach in our framework is straightforward. The government is the decision-maker, and the government's response is measured, for example, by levels of repression. The state of the world, θ, corresponds to the amount of antigovernment sentiment among citizens, and the social process, A, corresponds to the aggregate level of participation in antigovernment activities by citizens. In this case, we can think of the treatment as simply being the level of protest, so that $T(A) = A$.

The amount of protest is determined by the level of antigovernment sentiment (θ), the weather on the day of the protest (ω), and other idiosyncratic factors (ε). In this context, variation in weather corresponds to a contrast, (ω', ω''). The effect of interest is isolated by measuring the association between variation in the level of protest due to rainfall and variation in repression (or, in the case of the reduced-form relationship, the association between variation in rainfall and variation in repression). Many standard models in the literature on protest show how to micro-found protest in these kinds of concerns (Angeletos and Pavan 2013; Angeletos, Hellwig, and Pavan 2006, 2007; Bueno de Mesquita 2010; Casper and Tyson 2014; Edmond 2013; Shadmehr and Bernhardt 2011; Tyson and Smith 2018).

Empowering Female Candidates

Our second application corresponds to empirical studies that focus on the empowering effect of women officeholders by addressing whether one female candidate winning elected office causes other potential female candidates to run for office in subsequent elections (Baskaran and Hessami 2018; Broockman 2014; Ladam, Harden, and Windett 2018).

A victory by a female candidate could empower other female candidates through both direct and informational channels. A direct effect might arise if women prefer serving in office when there are other elected women who can, for example, serve as mentors or who may be desirable legislative collaborators. An informational effect might derive from potential female candidates interpreting earlier electoral victories by female candidates as evidence that voters are less discriminatory than they had previously suspected.

Identifying these causal effects is, of course, difficult because of a variety of endogeneity concerns—for instance, perhaps female candidates both win and emerge more often in more liberal districts. To address such concerns, a growing literature uses an election regression discontinuity design (Baskaran and Hessami 2018; Broockman 2014). The idea is to consider districts with male and female candidates running against one another. In some, the female candidate just barely wins, and in others, the male candidate just barely wins. If potential outcomes (i.e., future female candidates' willingness to run) are continuous at the electoral threshold, then comparing the frequency with which future female candidates run in districts that had a female vs. male victory provides an estimate of the causal effect of having a female candidate win on future female candidacies.

Again, we can represent this scenario within our framework. The decision-maker in our model corresponds to a future potential female candidate. We can think of her response function, r, as measuring her willingness to run. The state of the world in the model, θ, is the amount of antifemale bias among the electorate, and the outcome of the social process, A, corresponds to the vote differential between the female and male candidate in the initial election. It is determined by antifemale bias (θ), other features of the candidates or electoral environment that affect voting for the two candidates (ω), and idiosyncratic factors on election day (ε).

The treatment is the gender of the electoral winner in the initial election. This treatment is a monotone function of the vote differential, A, but it changes values

only once, at the electoral threshold. In particular, treatment can take one of two values—a female (F) or male (M) winner in the previous election:

$$T(A) = \begin{cases} F \text{ if } A \geq 0 \\ M \text{ if } A < 0. \end{cases}$$

The effect of interest is isolated by comparing districts where a female candidate just won against a male opponent to districts where a male candidate just won against a female opponent.

Indiscriminate Violence in Counterinsurgency

The final application we focus on concerns the effects of indiscriminate violence by counterinsurgents on rebel violence. Naturally, there are important empirical challenges that arise in isolating the effect of indiscriminate violence. The most common research design in the indiscriminate violence literature uses a difference-in-differences approach to compare matched localities in periods when they did or did not experience indiscriminate violence (Benmelech, Berrebi, and Klor 2014; Condra and Shapiro 2012; Lyall 2009). For instance, Lyall (2009) exploits a Russian doctrine of random shelling to generate variation in civilian casualties among similar villages and Condra and Shapiro (2012) match on levels of violence and exploit civilian deaths resulting from the violence.

Dell and Querubín (2017) take a somewhat different approach. They study the effect of aerial bombing by American forces in Vietnam on violent activity by the North Vietnamese insurgents (among other outcomes). The source of variation in their study comes from discontinuities in U.S. bombing strategy. Each locality was given a score based on a complex algorithm whose inputs included regular surveys of local commanders. Rounding thresholds in that algorithm led otherwise similar hamlets to experience different levels of bombing. We focus on Dell and Querubín's (2017) study, but return to compare their results with other studies later.

It is again straightforward to map this setting to our framework. The decision-maker in our model corresponds to North Vietnamese insurgent commanders, and the response function, r, measures the amount of violence engaged in by insurgents. The state of the world in the model, θ, might represent something like the level of U.S. resolve in the Vietnam War. The social process, A, corresponds to the level of bombing used by U.S. forces against a particular hamlet. It is determined by something like the level of U.S. commitment in Vietnam (θ), the dozens of factors that affect the hamlet's score in the bombing algorithm (ω), and idiosyncratic factors like data entry errors (ε).

The treatment is the amount of civilian casualties and destroyed infrastructure in the hamlet, $T(A)$, which is increasing in the amount of bombing. The effect of interest is isolated using a regression discontinuity at the rounding threshold, which compares hamlets that had essentially identical scores in the bombing algorithm but experienced very different levels of bombing.

RESULTS ON COMMENSURABILITY

We now return to our theoretical framework and consider the commensurability problem formally.

Our first result addresses commensurability for the total effect. The news is not good.

Theorem 1. *An experiment, $\mathcal{E}$, with associated information set $I^{\mathcal{E}}$, is commensurate with the ideal experiment for the total effect if and only if treatment status, T, is a sufficient statistic for $I^{\mathcal{E}}$ with respect to θ.*

This result says that an arbitrary experiment is only commensurate with the total effect in the ideal experiment if the decision-maker, whose response is the outcome of interest, has no information about the state of the world beyond what she learns from treatment status.

Suppose the decision-maker has additional information beyond that entailed in treatment status, for instance, because she observes the value of the shock (ω) or the outcome of the social process (A) directly. Then, for a fixed treatment, T, she will have a different posterior belief about θ in that experiment than she would have had in the ideal experiment. This means that the information effect is different in the two experiments, and when there are interactions between posterior beliefs and treatment status in the decision-maker's response (r), then the direct effects are also different. Thus, unless the difference in the information effects and the difference in the direct effects exactly off-set, the total effect in the nonideal experiment and the total effect in the ideal experiment are not the same. Theorem 1 confirms that, generically, these differences do not off-set, so that any nonideal experiment in which the decision-maker has information beyond knowledge of treatment status is not commensurate with the ideal experiment for the total effect. This commensurability problem matters because (short of estimating structural parameters) the total effect in the ideal experiment is often the quantity of most interest theoretically and substantively.

Although Theorem 1 shows that, in a setting where the decision-maker has information beyond knowledge of treatment status, a nonideal experiment does not estimate the total effect in the ideal experiment, perhaps it could estimate a different quantity of theoretical interest in the ideal experiment.

One intuition is that, if the decision-maker has enough extra information, perhaps there will be no information left in treatment assignment. If this were the case, maybe the nonideal experiment would estimate the direct effect of treatment in the ideal experiment, purged of any information effect. The next result shows that this intuition is correct, but only under an additional fairly stringent condition.

Theorem 2. *An experiment $\mathcal{E}$ with associated information set $I^{\mathcal{E}}$, where $I^{\mathcal{I}}$ is not sufficient for $I^{\mathcal{E}}$ with respect to θ, is commensurate with the ideal experiment for the direct effect if and only if*

1. *the information effect in $\mathcal{E}$, $\iota^{\mathcal{E}}(\omega', \omega'')$, is equal to zero for almost every contrast, (ω', ω'');*
2. *the decision-maker's response function, $r(T, \pi)$, is additively separable between its two arguments for almost*

every ω, i.e., if there exist functions $\rho^1 : \mathcal{T} \to \mathbb{R}$ and $\rho^2 : [0,1] \to \mathbb{R}$ such that for almost every ω:

$$\mathbb{E}_{\theta,\varepsilon}\left[r\left(T_{\theta,\varepsilon}^\omega, \pi(I(\theta,\omega,\varepsilon))\right)\right]$$
$$= \mathbb{E}_{\theta,\varepsilon}\left[\rho^1\left(T_{\theta,\varepsilon}^\omega\right)\right] + \mathbb{E}_{\theta,\varepsilon}\left[\rho^2(\pi(I(\theta,\omega,\varepsilon)))\right].$$

The first condition requires that there be no information in treatment status in the nonideal experiment. But even this is not enough. The second requirement is that the decision-maker's response be additively separable in the direct effect of treatment and her posterior beliefs about the state of the world. If a decision-maker's response to treatment depends on her beliefs, then even if there is no information effect in the experiment, the direct effects will be different between a nonideal and ideal experiment because the posterior beliefs will be different. That is, without additive separability, because the decision-maker's posterior beliefs are different in the two experiments, the same change in treatment status will have different effects on the decision-maker's response. The second condition, then, is closely tied to concerns about heterogeneous treatment effects, a topic we will return to when we discuss the application to female candidates in greater detail below.

COMMENSURABILITY IN THE APPLICATIONS

Motivated by the results in the previous section, we now ask whether some prominent research designs in political science and economics generate estimates of theoretically interpretable effects in their associated ideal experiments. This involves asking two questions.

First, motivated by Theorem 1, we ask whether prominent empirical approaches satisfy the requirement that treatment assignment is plausibly a sufficient statistic for the decision-maker's information set in the relevant empirical settings. We argue that in our first two applications discussed above—the efficacy of protest and the empowerment of female candidates—this is not the case. Thus, Theorem 1 indicates that in these applications, the research designs commonly used in empirical research are not estimating the total effect in the ideal experiment. By contrast, we argue that this condition is met in Dell and Querubín's (2017) study of indiscriminate bombing in Vietnam, and consequently, Theorem 1 suggests their study is estimating the total effect in the ideal experiment.

Second, motivated by Theorem 2, we ask whether the conditions for commensurability for the direct effect are met. We argue that the answer is again no for the literature on the efficacy of protest. But in our second application, on the empowerment of women candidates, we argue that at least the first condition (zero information effect in the nonideal experiment) is met. Thus, Theorem 2 highlights that commensurability for the direct effect comes down to answering a substantive question about the plausibility of the additive separability of the decision-maker's response.

Efficacy of Protest

The application of Theorem 1 to the empirical literature on the efficacy of political mobilization is straightforward. The ideal experiment for understanding the effect of protest behavior on government response involves a shock to protest behavior that is not observable by the government. But the actual research designs used in the literature exploit variation in protest behavior due to weather shocks. Such studies may be estimating perfectly sensible causal estimates; after all, weather shocks do manipulate protest size. But weather shocks are also observable to the government (that is, $\omega \in I^{\mathcal{E}}(\theta,\omega,\varepsilon)$). This affects how the government updates its beliefs in response to changes in protest behavior. As such, the estimates of the efficacy of protest obtained using weather shocks do not estimate the total effect of a change in protest behavior in the ideal experiment. Put differently, they do not estimate the effect of a change in protest behavior that is due to, say, actual changes to the underlying level of citizen discontent or unobservable changes in the capacity of antigovernment groups.

There is something at stake here. Many of the most prominent contributions in this literature are interested in the question of whether protests are productive or counterproductive. There may, of course, be competing effects. Suppose the direct effect of protests is counterproductive (e.g., governments get angry at protestors and want to repress them) but the informational effect is positive (e.g., when governments learn their citizens are truly angry, they may be inclined to change policy in ways that are beneficial to the citizens). The estimate from a research design that does not mimic the informational environment of the ideal experiment, and thus estimates a combination of a direct effect and an information effect that is infected by the government's interpretation of what the size of a protest means in light of the weather, and may get the sign of the total effect wrong relative to the ideal experiment. This could lead to incorrect, or at least difficult to interpret, conclusions about the efficacy of protest as a means to achieving political change by citizens.

Theorem 2 also implies that research designs using weather shocks do not estimate the direct effect of protest in the ideal experiment. To estimate the direct effect, there would need to be no information for the government from a change in protest size induced by weather. And, moreover, the way that the government responded to a disruptive protest would need to be additively separable from its assessment of its citizens' level of antigovernment sentiment.

The question of additive separability is, of course, a substantive one. However, it seems unlikely that a government's direct response to protestors in the streets is independent of its overall sense of public dissatisfaction.

We can provide a bit more formal guidance for whether there is zero information effect in the nonideal experiment where the decision-maker observes the shock. Let ϕ be the (differentiable) density function of idiosyncratic noise, ε. Then, assuming the decision-maker updates her beliefs using Bayes' rule, her

posterior belief that $\theta = \bar{\theta}$ at a level of protest A and a shock ω (with both observable) is

$$\pi(\{A, \omega\})$$
$$= \frac{p\phi\left(F^{-1}(A) - \bar{\theta} - \omega\right)}{p\phi(F^{-1}(A) - \bar{\theta} - \omega) + (1-p)\phi(F^{-1}(A) - \underline{\theta} - \omega)}.$$

For the information effect to be zero, for almost every A, this posterior belief needs to be constant in ω almost everywhere. Differentiating, a zero information effect requires that for almost all ω:

$$\frac{\phi'}{\phi}\left(F^{-1}(A) - \bar{\theta} - \omega\right) = \frac{\phi'}{\phi}\left(F^{-1}(A) - \underline{\theta} - \omega\right).$$

This condition holds if and only if ϕ is almost everywhere log-linear, a highly restrictive condition.

We want to emphasize that we do not intend these results as critiques of the specific papers using weather shocks to estimate the efficacy of protests. Seen from the perspective of the project of doing better causal inference, those papers make important advances. Rather, our goal is to shed some light on whether we can interpret those estimates in terms of important theoretical quantities in the ideal experiment to highlight the conceptual difficulties associated with studying the effect of behavior on behavior.

Empowering Female Candidates

The best identified empirical studies on the empowerment of female candidates use a regression discontinuity design (Baskaran and Hessami 2018; Broockman 2014).

In the ideal experiment, the decision-maker (a future potential female candidate) observes only treatment status—i.e., whether a male or female won the last election. Hence, in the ideal experiment, the decision-maker has posterior beliefs:

$$\pi\left(I^{\mathcal{I}}(\theta, \omega, \varepsilon)\right) = \begin{cases} \Pr\left(\theta = \bar{\theta} | A^{\omega}_{\theta,\varepsilon} \geq 0\right) \text{ if } T^{\omega}_{\theta,\varepsilon} = F \\ \Pr\left(\theta = \bar{\theta} | A^{\omega}_{\theta,\varepsilon} < 0\right) \text{ if } T^{\omega}_{\theta,\varepsilon} = M. \end{cases}$$

But now consider the actual electoral regression discontinuity design used to assess the effect of a female victory on future female candidacy. Presumably, potential future female candidates are paying some attention to nearby races precisely because such races are informative about their own electoral prospects. Hence, it seems likely that she observes not just who won the previous election (T) but the vote totals in that election as well (A). That is, $A^{\omega}_{\theta,\varepsilon} \in I^{\mathcal{E}}(\theta, \omega, \varepsilon)$.

Notice, the binary variable T (whether a female candidate wins) is entirely determined by the continuous variable A (the vote differential). So treatment status is certainly not a sufficient statistic for the decision-maker's information with respect to θ (voter bias). Indeed, instead, A is a sufficient statistic for T. That is, the decision-maker can ignore who won in forming her posterior beliefs; all she needs to know is the vote differential. Who won still matters for her action, but only because of direct effects. Thus, even before discussing any of the specifics of the regression

discontinuity, we already know by Theorem 1 that the actual research design is not commensurate with the ideal experiment for the total effect.

Nonetheless, it will be instructive to work through what does happen in an experiment that represents the actual research design, so we can think more carefully about the regression discontinuity design and the direct effect in the ideal experiment. Given that she observes the vote total from the previous election, a Bayesian decision-maker has posterior beliefs:

$$\pi\left(I^{\mathcal{E}}(\theta, \omega, \varepsilon)\right) = \Pr\left(\theta = \bar{\theta} | A^{\omega}_{\theta,\varepsilon}\right).$$

The regression discontinuity design, of course, estimates behavior at the electoral threshold $A = 0$. In this limit, the decision-maker's posterior belief goes to

$$\Pr\left(\theta = \bar{\theta} | A^{\omega}_{\theta,\varepsilon} = 0\right).$$

This posterior belief is never equal to the posterior belief in the ideal experiment. That is, T is not sufficient for (T, A), and hence by Theorem 1, the research design does not estimate the total effect in the ideal experiment.

As before, this is not a critique of the research design. In our model, continuity of potential outcomes at the threshold is satisfied. Rather, the problem is that future potential candidates know the electoral returns from the prior election, so they have more information than the decision-maker in the ideal experiment.

The fact that, when vote totals are observed, posterior beliefs are constant in treatment in the RD (i.e., in the limit beliefs are always $\Pr(\theta = \bar{\theta} | A = 0)$) raises the possibility that, despite not estimating the total effect in the ideal experiment, the election RD may isolate the direct effect. In particular, the electoral RD satisfies the first requirement of Theorem 2—the information effect in the nonideal experiment that represents the research design is zero because beliefs are constant in treatment.

Theorem 2 then says that whether the election RD identifies the direct effect of female victory in the ideal experiment comes down to a substantive question—whether female candidates' willingness to run is an additively separable function of treatment status (whether a woman won) and posterior beliefs about antifemale bias in the electorate. This question, put another way, asks whether the direct effect of having a female candidate win on another woman's willingness to run is different depending on that potential candidate's beliefs about the amount of antifemale bias in the electorate. There are reasonable arguments in either direction. One possibility is that it is simply more productive or rewarding to serve in office when there are other women in office, regardless of what voters think. This would suggest that additive separability is plausible. Another possibility is that, say, when voters are particularly biased, the sense of solidarity associated with another politician who has also overcome considerable bias is heightened, suggesting that additive separability does not hold. We don't take a stand regarding which assumption is right. Rather, we simply want to emphasize that Theorem 2 highlights this substantive question, which one must answer to know

whether the election RD identifies a quantity of theoretical interest in the ideal experiment—namely, the direct effect of a female victory on future female candidates' willingness to run.

This discussion of additive separability raises a final point about Theorem 2 regarding heterogeneous effects. There is no question that when the decision-maker observes vote shares, the regression discontinuity design yields an unbiased estimate of a local average direct effect at the electoral threshold—potential outcomes are continuous and beliefs are constant at the threshold. This is true whether or not the decision-maker's response satisfies additive separability. Yet, without additive separability, this effect is not commensurate with the direct effect at the same electoral threshold in the ideal experiment.

This is because, when additive separability fails, there are heterogeneous direct effects across different beliefs. But the commensurability problem here is not the standard concern that local average treatment effects need not be the same as the average treatment effect when effects are heterogenous. Here, we are comparing two local direct effects (in the nonideal and ideal experiments), holding the localness constant—in both cases, we are talking about the direct effect at the same electoral threshold. The commensurability problem comes from the fact that, without additive separability, the local direct effects at that same threshold are different quantities in the nonideal and ideal experiments because the decision-maker's beliefs are different.

These theoretical observations are, we think, particularly salient in light of recent evidence from election RD studies that finds null effects, suggesting that female victories do not empower future female candidates (Broockman 2014). The implication of Theorem 1 is that we should be careful about overinterpreting such a finding. The election RD purges the estimate of any informational effect of a female victory. Thus, at most, what we can conclude from such a finding is that there is no evidence of a direct effect of a female victory on empowerment of future female candidates. There still might be an informational effect that is washed out by the research design. Theorem 2 says that we should even be hesitant about that conclusion if we do not believe that the direct and informational components of a potential candidate's decision are additively separable.

Indiscriminate Violence in Counterinsurgency

The application of Theorem 1 to Dell and Querubín's (2017) study of indiscriminate bombing in Vietnam leads to different conclusions than in the previous applications we consider. Recall, in relating our model to the empirical setting, we said that ω represents the inputs to the bombing algorithm scores, and contrasts correspond to small changes to score inputs that got magnified by rounding. Those scores were created by an incredibly complex process that took a huge amount of subjective data and returned simple indexes. The rounding built into that process was unobservable to the insurgents (and, indeed, even to the Americans). Hence, Theorem 1 implies that the effect estimated by

Dell and Querubín (2017) is commensurate with the total effect of bombing in the ideal experiment—treatment assignment is a sufficient statistic for the insurgents' information relative to the state of world, which represents U.S. resolve or strategy in Vietnam. As such, the effect of variation in bombing due to score rounding should have the same effect on the insurgents' beliefs, and thus response, as similar variation in bombing due to actual changes in America's underlying resolve or strategy.

Something similar can be argued for the effect identified by Lyall (2009) using random shelling by Russian artillery. Assuming Chechen rebels were unaware of the Russian military doctrine that called for random shelling, the effect of increased shelling in that setting is similar to the effect of increased bombing due to rounding errors in Vietnam. By contrast, the effect identified by Condra and Shapiro (2012) is quite different. In particular, their identification strategy exploits random collateral damage, holding fixed the amount of counterinsurgent activity. In that setting, if insurgents are aware of the amount of counterinsurgent activity, then it seems there is no extra informational content in whether a civilian happened to be killed. This, then, appears more similar to studies using the election RD to study female empowerment. The research design purges information effects. Hence, Condra and Shapiro's (2012) estimates are not commensurate with the total effect, but by Theorem 2, they may be commensurate with the direct effect if the direct and informational effects of civilian casualties on insurgency are additively separable.

CONCLUSION

We articulate the *commensurability problem*, which addresses the relationship between theoretical concepts and the empirical quantities that evaluate and confront theories. Specifically, it concerns when actual research designs estimate quantities of theoretical interest in ideal experiments.

We study this problem in a formal setting where treatment is determined by a process that itself has informational content that matters for outcomes. We establish two results. First, an experiment is only commensurate with the ideal experiment for the total effect if treatment status is a sufficient statistic for the decision-makers' information. When this is not the case, information effects are systematically different across nonideal and ideal experiments. Second, a nonideal experiment is commensurate with the ideal experiment for the direct effect if two conditions hold:

1. There is no information effect in the nonideal experiment.
2. The decision-maker's response function is additively separable in treatment and information.

These results apply to any empirical setting that exhibits both direct and informational channels. We apply our results to three substantive literatures: the efficacy of protest, the empowerment of female candidates, and indiscriminate violence in counterinsurgency. Doing

so shows how thinking about commensurability can help us better conceptualize the relationship between theoretical quantities of interest, ideal experiments, and the estimates generated by actual research designs. In particular, each of these literatures has made significant progress in terms of causal identification. However, our results suggest that this alone does not guarantee that they are estimating theoretically meaningful quantities from the ideal experiment. Finally, although we explicitly illustrate how our results apply to quasi-experimental settings, it is important to stress that the commensurability problem can arise in any empirical setting exhibiting both direct and informational channels. Such channels can certainly arise in experimental, as well as quasi-experimental settings. For instance, Grimmer, Messing, and Westwood (2012) experimentally study how congressional press releases impact voter approval, where the effect of congressional press releases might operate both through a direct psychological mechanism as well as through an informational mechanism where the press release is informative about the representative's ability to procure distributive spending.

Our results, of course, can be viewed as largely negative—showing ways in which even well-identified estimates may fail to capture quantities of theoretical interest. But we hope they can also be understood as constructively offering guidance for how to think about the design of empirical studies that can speak to theory when the treatments and outcomes of interest are behavior and information effects are important. In this spirit, our first result can be understood as encouraging empirical researchers interested in engaging theory to look for research designs where the variation in treatment behavior cannot be observed by the outcome agent. Dell and Querubín (2017) provide a nice example of when this is possible. And our second result can be understood as delineating the conditions under which the direct effect of treatment can be recovered by a research design and clarifying the key substantive questions that need to be answered to know whether this has been done successfully.

Our hope, then, is that this work will help with ongoing attempts to bridge the gap between careful research design for causal inference and the theoretical interpretability of the estimates obtained by such studies.

REFERENCES

Angeletos, George-Marios, and Alessandro Pavan. 2013. "Selection-Free Predictions in Global Games with Endogenous Information and Multiple Equilibria." *Theoretical Economics* 8 (3): 883–938.
Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan. 2006. "Signaling in a Global Game: Coordination and Policy Traps." *Journal of Political Economy* 114 (3): 452–84.
Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan. 2007. "Dynamic Global Games of Regime Change: Learning, Multiplicity, and the Timing of Attacks." *Econometrica* 75 (3): 711–56.
Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.
Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ashworth, Scott, and Ethan Bueno de Mesquita. 2014. "Is Voter Competence Good for Voters? Information, Rationality, and Democratic Performance." *American Political Science Review* 108 (3): 565–87.
Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2018. "Learning about Voter Rationality." *American Journal of Political Science* 32 (1): 37–54.
Banerjee, Abhijit V., Sylvain Chassang, and Erik Snowberg. 2017. "Decision Theoretic Approaches to Experiment Design and External Validity." In *Handbook of Economic Field Experiments*. Vol. 1, eds. Abhijit Vinayak Banerjee and Esther Duflo. Elsevier, 141–74.
Baskaran, Thushyanthan, and Zohal Hessami. 2018. "Does the Election of a Female Leader Clear the Way for More Women in Politics?" *American Economic Journal: Economic Policy* 10 (3): 95–121.
Benmelech, Efraim, Claude Berrebi, and Esteban F. Klor. 2014. "Counter-Suicide-Terrorism: Evidence from House Demolitions." *The Journal of Politics* 77 (1): 27–43.
Blackwell, David. 1951. "Comparison of Experiments." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley, CA: University of California Press, 93–102.
Broockman, David E. 2014. "Do Female Politicians Empower Women to Vote or Run for Office? A Regression Discontinuity Approach." *Electoral Studies* 34: 190–204.
Bueno de Mesquita, Ethan. 2010. "Regime Change and Revolutionary Entrepreneurs." *American Political Science Review* 104 (3): 446–65.
Casper, Brett Allen, and Scott A. Tyson. 2014. "Popular Protest and Elite Coordination in a Coup D'etat." *The Journal of Politics* 76 (2): 548–64.
Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *The American Economic Review* 102 (4): 1279–309.
Collins, William J., and Robert A. Margo. 2007. "The Economic Aftermath of the 1960s Riots in American Cities: Evidence from Property Values." *The Journal of Economic History* 67 (4): 849.
Condra, Luke N., and Jacob N. Shapiro. 2012. "Who Takes the Blame? The Strategic Effects of Collateral Damage." *American Journal of Political Science* 56 (1): 167–87.
Conlisk, John. 1973. "Choice of Response Functional Form in Designing Subsidy Experiments." *Econometrica* 41 (4): 643.
Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.
Dell, Melissa. 2012. "Path Dependence in Development: Evidence from the Mexican Revolution." Harvard typescript.
Dell, Melissa, and Pablo Querubín. 2017. "Nation Building through Foreign Intervention: Evidence from Discontinuities in Military Strategies." *Quarterly Journal of Economics* 133 (2): 701–64.
Edmond, Chris. 2013. "Information Manipulation, Coordination, and Regime Change." *The Review of Economic Studies* 80 (4): 1422–58.
Eggers, Andrew C. 2017. "Quality-Based Explanations of Incumbency Effects." *The Journal of Politics* 79 (4): 1315–28.
Feyerabend, Paul K. 1962. "Explanation, Reduction and Empiricism." In *Scientific Explanation, Space, and Time (Minnesota Studies in the Philosophy of Science, Volume III)*, eds. Herbert Feigl and Grover Maxwell. Minneapolis: University of Minnesota Press, 28–97.
Fowler, Anthony. 2018. "A Bayesian Explanation for the Effect of Incumbency." *Electoral Studies* 53: 66–78.
Fudenberg, Drew, and David K. Levine. 2019. "Learning in Games and the Interpretation of Natural Experiments". MIT typescript.
Goldberger, Arthur S. 1972. "Structural Equation Methods in the Social Sciences." *Econometrica: Journal of the Econometric Society* 40 (6): 979–1001.
Gould, Eric D., and Esteban Klor. 2010. "Does Terrorism Work?" *Quarterly Journal of Economics* 125 (4): 1459–510.
Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2012. "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation." *American Political Science Review* 106 (4): 703–19.
Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica: Journal of the Econometric Society* 12: iii–115.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115 (1): 45–97.

Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156 (1): 27–37.

Henderson, John, and John Brooks. 2016. "Mediating the Electoral Connection: The Information Effects of Voter Signals on Legislative Behavior." *The Journal of Politics* 78 (3): 653–69.

Hendrix, Cullen S., and Idean Salehyan. 2012. "Climate Change, Rainfall, and Social Conflict in Africa." *Journal of Peace Research* 49 (1): 35–50.

Huet-Vaughn, Emiliano. 2013. "Quiet Riot: The Causal Effect of Protest Violence." UC Berkeley typescript.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Izzo, Federical, Torun Dewan, and Stephane Wolton. 2018. "Cumulative Knowledge in the Social Sciences: The Case of Improving Voters' Information." Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3239047.

Jaeger, David A., Esteban F. Klor, Sami H. Miaari, and M.Daniele Paserman. 2012. "The Struggle for Palestinian Hearts and Minds: Violence and Public Opinion in the Second Intifada." *Journal of Public Economics* 96 (3–4): 354–68.

Koopmans, Tjalling C., and Olav Reiersol. 1950. "The Identification of Structural Characteristics." *The Annals of Mathematical Statistics* 21 (2): 165–81.

Kreps, David M. 2012. *Microeconomic Foundations I: Choice and Competitive Markets*. Princeton, NJ: Princeton University Press.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Ladam, Christina, Jeffrey J. Harden, and Jason H. Windett. 2018. "Prominent Role Models: High-Profile Female Politicians and the Emergence of Women as Candidates for Public Office." *American Journal of Political Science* 62 (2): 369–81.

Lyall, Jason. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya." *Journal of Conflict Resolution* 53 (3): 331–62.

Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott. 2013. "Do Political Protests Matter? Evidence from the Tea Party Movement." Harvard University typescript.

Mas-Collel, Andreu, Michael D. Whiston, and Jerry R. Green. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Milnor, John Willard. 1997. *Topology from the Differentiable Viewpoint*. Princeton, NJ: Princeton University Press.

Prato, Carlo, and Stephane Wolton. 2018. "Electoral Imbalances and Their Consequences." *The Journal of Politics* 80 (4): 1168–82.

Ritter, Emily Hencken, and Cortenay R. Conrad. 2016. "Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression." *American Political Science Review* 110 (1): 85–99.

Sarsons, Heather. 2015. "Rainfall and Conflict: A Cautionary Tale." *Journal of Eevelopment Economics* 115: 62–72.

Shadmehr, Mehdi, and Dan Bernhardt. 2011. "Collective Action with Uncertain Payoffs: Coordination, Public Signals, and Punishment Dilemmas." *American Political Science Review* 105: 829–51.

Shapiro, Jacob N., and Nils B. Weidmann. 2015. "Is the Phone Mightier Than the Sword? Cellphones and Insurgent Violence in Iraq." *International Organization* 69 (2): 247–74.

Sun, Jessica, and Scott A. Tyson. 2019. "Theoretical Implications of Empirical Models: An Application to Conflict Studies." Available at SSRN 3258725.

Tyson, Scott A., and Alastair Smith. 2018. "Dual-Layered Coordination and Political Instability: Repression, Co-optation, and the Role of Information." *The Journal of Politics* 80 (1): 44–58.

Wolton, Stephane. 2019. "Are Biased Media Bad for Democracy?" *American Journal of Political Science* 63 (3): 548–562.

APPENDIX

Before presenting our main results, we first establish the following Lemma:

Lemma 1 *Let $\theta \in \{\underline{\theta}, \bar{\theta}\}$ be Bernoulli distributed with parameter $p = P(\theta = \bar{\theta})$ and ε follow distribution function Ψ with support a subset of $\mathbb{R}$. Fix $c \in \mathbb{R}$ and let X be a compact Borel-measurable subset of $\mathbb{R}$. If $f(x, p; \theta, \varepsilon)$ is a smooth function whose gradient, $\nabla f(x, p; \theta, \varepsilon) : X \times [0,1] \times \{\underline{\theta}, \bar{\theta}\} \times \mathbb{R} \to \mathbb{R}^4$, has full rank and where*

$$f(x, p; \theta, \varepsilon) \neq c$$

for almost every $(x, p, \theta, \varepsilon)$, then

$$F(x; p) \equiv \mathbb{E}_\varepsilon[pf(x, p; \bar{\theta}, \varepsilon) + (1-p)f(x, p; \underline{\theta}, \varepsilon); p] \neq c,$$

for almost all values of $x \in X$ and $p \in [0, 1]$.

Proof: For a fixed constant, c, let

$$\mathcal{P} = \{(x, p) \mid F(x; p) = c\} \subset X \times [0, 1],$$

and define the projection for a fixed x as

$$\bar{\mathcal{P}}_x = \{p \mid F(x; p) = c\} \subset [0, 1].$$

If the set $\mathcal{P}$ is empty, then we are done, so suppose instead that $\mathcal{P}$ is nonempty. Fix x and define the function

$$\hat{F}(x; \sigma, p) \equiv \mathbb{E}_\varepsilon[pf(x, \sigma; \bar{\theta}, \varepsilon) + (1-p)f(x, \sigma; \underline{\theta}, \varepsilon); p].$$

Next, for a fixed x and c, define the function $\sigma : [0, 1] \to [0, 1]$:

$$\hat{F}(x; \sigma(p), p) = c,$$

and note that σ is smooth.

Because $\hat{F}$ coincides with F only at fixed points of σ, any element $q \in \bar{\mathcal{P}}_x$ requires that $\sigma(q) = q$, and thus, in any open neighborhood of q contained in $\bar{\mathcal{P}}_x$, it must be that

$$\sigma'(q) = 1.$$

Because q was arbitrary, this implies that $\bar{\mathcal{P}}_x$ is contained in the set of critical points of σ. Then, by Sard's Theorem (Milnor 1997, 10), $\sigma(\overline{\mathcal{P}_x})$ is a measure zero subset of $[0, 1]$. Because x was arbitrary, this is true for any $x \in X$, and thus, the set $\mathcal{P}$ is a measure zero subset of $X \times [0, 1]$ according to the product measure. ∎

Proof of Theorem 1: We are interested in assessing when the total effect in the experiment $\mathcal{E}, \tau^{\mathcal{E}}$, equals the total effect in the ideal experiment, $\tau^{\mathcal{I}}$, at almost every pair of contrasts and prior, $\langle (\omega', \omega''), p \rangle$. Because the effect function comprised a difference in responses, by the Mean Value Theorem,

$$\tau^{\mathcal{I}}(\omega', \omega'') = \tau^{\mathcal{E}}(\omega', \omega'')$$

for almost every $\langle (\omega', \omega''), p \rangle$ if and only if for some constant $c \in \mathbb{R}$,

$$\mathbb{E}_{\theta,\varepsilon}\left[r\left(T_{\theta,\varepsilon}^\omega, \pi\big(I^{\mathcal{I}}(\theta, \omega, \varepsilon)\big) \right) - r\left(T_{\theta,\varepsilon}^\omega, \pi\big(I^{\mathcal{E}}(\theta, \omega, \varepsilon)\big) \right) \right] = c,$$

(7)

for almost every (ω, p).

First, we set up our argument by defining three functions. Second, we use these functions to show that an experiment $\mathcal{E}$ is commensurate with the ideal experiment for the total effect τ if and only if a particular condition holds for almost every shock. Finally, we show that the subset of shocks on which that condition holds has Lebesgue measure zero on the set of shocks.

For an experiment, $\mathcal{E}$, and a fixed (θ, ε), define the function $g_{\theta,\varepsilon}(\omega) : \Omega \to \mathbb{R}$ by

$$g_{\theta,\varepsilon}^{\mathcal{E}}(\omega) = \pi\big(I^{\mathcal{E}}(\theta,\omega,\varepsilon)\big) - \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big),$$

which measures the difference between the decision-maker's posterior belief in the ideal experiment and the decision-maker's posterior belief in an experiment $\mathcal{E}$.

We can rewrite Condition (7) as

$$\mathbb{E}_{\theta,\varepsilon}\Big[r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big)\Big)$$
$$-r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big) + g_{\theta,\varepsilon}^{\mathcal{E}}(\omega)\Big)\Big] = c \quad,$$

for some constant $c \in \mathbb{R}$ and almost every (ω, p).

Because r is one-to-one and smooth, its derivative has full rank. So, by Lemma 1, commensurability for the total effect requires

$$r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big)\Big)$$
$$- r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big) + g_{\theta,\varepsilon}^{\mathcal{E}}(\omega)\Big) = c$$

for some constant $c \in \mathbb{R}$ and almost every (ω, p).

Define the function $\Delta : \Omega \times [-1,1] \to \mathbb{R}$:

$$\Delta_{\theta,\varepsilon}(\omega, z) = r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big)\Big)$$
$$- r\Big(T_{\theta,\varepsilon}^{\omega}, \pi\big(I^{\mathcal{I}}(\theta,\omega,\varepsilon)\big) + z\Big).$$

Because r is one-to-one and smooth, its derivative has full rank, and this implies that Δ is smooth and its derivative has full rank.

Now, define the implicit function $z_{\theta,\varepsilon}(\omega,c) : \Omega \times \mathbb{R} \to [-1,1]$ as

$$\Delta_{\theta,\varepsilon}(\omega, z_{\theta,\varepsilon}(\omega,c)) = c.$$

This is the mapping that keeps the value of Δ constant at some value c for every ω. Notice, because the derivative of Δ has full rank, the Implicit Function Theorem implies that the derivative of z has full rank.

So the experiment $\mathcal{E}$ is commensurate with the ideal experiment for the total effect if and only if there is some $c \in \mathbb{R}$ such that

$$z_{\theta,\varepsilon}(\omega,c) = g_{\theta,\varepsilon}^{\mathcal{E}}(\omega), \tag{8}$$

for almost all $(\omega, p) \in \Omega \times (0,1)$. Our strategy is to show that the set of shocks for which equation (8) holds has Lebesgue measure zero, which will imply that the set of contrasts and priors for which equation (8) holds, has measure zero under the product measure.

Using equation (8), implicitly define the function $H : \Omega \times \mathbb{R} \to \Omega$ by

$$z_{\theta,\varepsilon}(H(\omega,c),c) = g_{\theta,\varepsilon}^{\mathcal{E}}(\omega),$$

which, by the Implicit Function Theorem, is a continuously differentiable function whose derivative has full rank. The experiment $\mathcal{E}$ is commensurate with the ideal experiment if and only if H is the identity function almost everywhere.

For sufficiency, suppose T is a sufficient statistic for $I^{\mathcal{E}}$ with respect to θ, implying that the decision-maker's posterior belief in the ideal experiment is equal to her posterior belief in the experiment $\mathcal{E}$ almost everywhere. When this is the case, two conditions are satisfied. First, because the posterior in the ideal experiment and the posterior in experiment $\mathcal{E}$ are the same,

$$g_{\theta,\varepsilon}^{\mathcal{E}}(\omega) = 0,$$

for every $\omega \in \Omega$. Second, again because the posterior belief in $\mathcal{E}$ is the same as the posterior belief in the ideal experiment,

$$\Delta_{\theta,\varepsilon}(\omega,0) = 0,$$

for every $(\theta, \varepsilon, \omega)$, completing the first part of the argument.

For necessity, suppose that T is not a sufficient statistic for $I^{\mathcal{E}}$ with respect to θ, implying that the decision-maker's posterior belief in the ideal experiment is different from her posterior belief in the experiment $\mathcal{E}$, on a set of $(\theta, \omega, \varepsilon)$ of positive measure. This implies that $g_{\theta,\varepsilon}^{\mathcal{E}}(\omega) \neq 0$ on a set of $(\theta, \omega, \varepsilon)$ of positive measure; call this set Γ.

Because Γ has positive measure, for commensurability, $\mathcal{E}$ must capture the total effect in $\mathcal{I}$ almost everywhere on Γ despite the fact that the posterior beliefs are different in the two experiments. We will show that, instead, $\mathcal{E}$ captures the total effect in $\mathcal{I}$ on at most a measure zero subset of Γ.

Denote the ω-dimension of Γ by $\bar{\Omega}$. Denote the set of shocks that lead the total effect in $\mathcal{E}$ and in $\mathcal{I}$ to be equal when $g_{\theta,\varepsilon}^{\mathcal{E}}(\omega) \neq 0$ by

$$\mathcal{W}^{\mathcal{E}}(\Omega,c) = \Big\{\omega \mid z_{\theta,\varepsilon}(\omega,c) = g_{\theta,\varepsilon}^{\mathcal{E}}(\omega)\Big\} \subset \bar{\Omega}.$$

Our strategy is to show that the set $\mathcal{W}^{\mathcal{E}}$ has Lebesgue measure zero in $\bar{\Omega} \times \{c\}$.

If $\mathcal{W}^{\mathcal{E}}$ is empty, the result is immediate, and so we proceed under the supposition that $\mathcal{W}^{\mathcal{E}}$ is nonempty. Consider an arbitrary shock $\omega \in \mathcal{W}^{\mathcal{E}}$, and let $N_{\omega} \subset \bar{\Omega}$ be an open neighborhood of ω. Because the function H is smooth, for any other shock in $N_{\omega} \cap \mathcal{W}^{\mathcal{E}}$, the first derivative (Jacobian) of H must satisfy

$$\frac{\partial H(\omega,c)}{\partial \omega} = 1, \tag{9}$$

almost everywhere. This implies that the set of shocks in $N_{\omega} \cap \mathcal{W}^{\mathcal{E}}$ is not of maximal rank and, hence, is contained in the set of critical points of the smooth mapping H. Because the choice of ω was arbitrary, this implies that the whole set $\mathcal{W}^{\mathcal{E}}$ is contained in the set of critical points of H. Because H is a smooth function whose derivative has full rank, Sard's Theorem (Milnor 1997, 10) establishes that the set $H(\mathcal{W}^{\mathcal{E}},c)$ has Lebesgue measure 0 in $\bar{\Omega} \times \{c\}$. Because H is the identify function on $\mathcal{W}^{\mathcal{E}}$, this establishes that $\mathcal{W}^{\mathcal{E}}$ has Lebesgue measure 0 on $\bar{\Omega}$. Hence, the set of $(\omega,p) \subset \bar{\Omega} \times (0,1)$ for which $\mathcal{E}$ is commensurate with the ideal experiment for the total effect has measure zero under the product measure. ∎

Proof of Theorem 2: Fix an information set I. If r is an additively separable function at almost every ω, then there exist functions $\rho^1 : \mathcal{T} \to \mathbb{R}$ and $\rho^2 : [0,1] \to \mathbb{R}$ such that for almost every ω, we can write

$$\mathbb{E}_{\theta,\varepsilon}\Big[r\Big(T_{\theta,\varepsilon}^{\omega}, \pi(I(\theta,\omega,\varepsilon))\Big)\Big]$$
$$= \mathbb{E}_{\theta,\varepsilon}\Big[\rho^1\Big(T_{\theta,\varepsilon}^{\omega}\Big) + \rho^2(\pi(I(\theta,\omega,\varepsilon)))\Big]. \tag{10}$$

This implies that in an experiment with associated information set I, for almost every contrast (ω', ω''), the total effect at (ω', ω'') is

$$\tau^{\mathcal{E}}(\omega',\omega'') = \frac{\mathbb{E}_{\theta,\varepsilon}\left[\overbrace{\rho^1\left(T_{\theta,\varepsilon}^{\omega''}\right) - \rho^1\left(T_{\theta,\varepsilon}^{\omega'}\right)}^{\text{direct effect}} + \overbrace{\rho^2(\pi(I(\theta,\omega'',\varepsilon))) - \rho^2(\pi(I(\theta,\omega',\varepsilon)))}^{\text{information effect}}\right]}{\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right]}.$$

Recall from Definition 7 that an experiment $\mathcal{E}$ is commensurate with the ideal experiment for the direct effect if the total effect in the experiment $\mathcal{E}$ almost always captures the direct effect in the ideal experiment. This is only true if

$$\tau^{\mathcal{E}}(\omega',\omega'') = \frac{\mathbb{E}_{\theta,\varepsilon}\left[\rho^1\left(T_{\theta,\varepsilon}^{\omega''}\right) - \rho^1\left(T_{\theta,\varepsilon}^{\omega'}\right)\right]}{\mathbb{E}_{\theta,\varepsilon}\left[T_{\theta,\varepsilon}^{\omega''} - T_{\theta,\varepsilon}^{\omega'}\right]}, \qquad \textbf{(11)}$$

at almost every contrast and prior $\langle(\omega',\omega''),p\rangle \in \mathcal{C}\times(0,1)$.

For sufficiency, note that if the information effect is zero and there is additive separability of the response function, then Condition (11) holds almost everywhere.

Now consider necessity. We proceed in two steps. The first step addresses the information effect, and the second step addresses additive separability of the response function. For both steps, we proceed by contradiction.

Part 1: Suppose that r is additively separable for almost all $(\theta,\varepsilon,\omega)$. Suppose further that the information effect in the experiment $\mathcal{E}$ is not zero on a set of contrasts with positive measure, denoted by $\mathcal{K}$. Here, we will show the necessity of having a zero information effect. Condition (11) does not hold on the set $\mathcal{K}$. Thus, the total effect in experiment $\mathcal{E}$ does not capture the direct effect in the ideal experiment on $\mathcal{K}$. Because the set $\mathcal{K}$ is of positive measure, experiment $\mathcal{E}$ is not commensurate with the ideal experiment for the direct effect.

Part 2: We now consider additive separability, supposing that the information effect in the experiment $\mathcal{E}$ is zero at almost all contrasts (ω',ω''). We proceed by contradiction and suppose that r is not additively separable on a set of positive measure and focus on this set.[8]

Define the function $L : \mathcal{C} \to \mathbb{R}$, by

$$L^{\mathcal{E}}(\omega',\omega) = r\left(T_{\theta,\varepsilon}^{\omega'}, \pi\left(I^{\mathcal{E}}(\theta,\omega,\varepsilon)\right)\right).$$

Note that this is a smooth function whose derivative has full rank.

Because the response function is smooth, a zero information effect for almost every contrast is equivalent to

$$\mathbb{E}_{\theta,\varepsilon}\left[\frac{\partial L^{\mathcal{E}}(\omega',\omega)}{\partial \omega}\right]$$
$$= \mathbb{E}_{\theta,\varepsilon}\left[\frac{\partial r\left(T_{\theta,\varepsilon}^{\omega'}, \pi\left(I^{\mathcal{E}}(\theta,\omega,\varepsilon)\right)\right)}{\partial \pi} \cdot \frac{\partial \pi\left(I^{\mathcal{E}}(\theta,\omega,\varepsilon)\right)}{\partial \omega}\right] = 0,$$

$$\textbf{(12)}$$

almost everywhere.[9] Equation (12) is a linear ordinary differential equation whose solution is an arbitrary function $L^*(\omega')$, which is independent of ω.[10] That there are no other kinds of solutions follows from the smoothness of L and the Picard–Lindelöf Theorem.

There are three ways for (12) to hold for almost every ω:

(i) Results from taking expectations over (θ,ε);
(ii) Additive separability of the response function, r;
(iii) The posterior belief $\pi\left(I^{\mathcal{E}}(\theta,\omega,\varepsilon)\right)$ is constant in ω for almost every $(\theta,\omega,\varepsilon)$.

The first is ruled out by Lemma 1. The second is ruled out by hypothesis. So, suppose that r is not additively separable, but the posterior is constant in ω for almost every $(\theta,\varepsilon,\omega)$. Using arguments similar to those in the proof of Theorem 1, by Lemma 1 and the mean value theorem, commensurabiltiy for the direct effect holds if and only if

$$\frac{\partial L^{\mathcal{E}}(\omega',\omega)}{\partial \omega'} = \frac{\partial L^{\mathcal{I}}(\omega',\omega)}{\partial \omega'}.$$

To see that this is nongeneric, define the function $y \colon \Omega \to \Omega$ by

$$\frac{\partial L^{\mathcal{E}}(\omega',\omega)}{\partial \omega'} = \frac{\partial L^{\mathcal{I}}(\omega',y)}{\partial \omega'},$$

and note that the right-hand side corresponds to the direct effect in the ideal experiment if and only if y is the identity function. Following an identical argument as in the proof of Theorem 1, Sard's Theorem again implies that this holds on a set of at most measure zero. ∎

---

[8] This rules out functions of the form (10), which includes linear functions.

[9] This follows from the definition above by varying only ω'.

[10] That (12) is an ordinary differential equation follows because the derivative of the first argument does not appear.