European Journal for Philosophy of Science (2021) 11: 42
https://doi.org/10.1007/s13194-021-00350-z

PAPER IN THE PHILOSOPHY OF THE SOCIAL SCIENCES AND HUMANITIES

Reactivity in social scientific experiments: what is it and how is it different (and worse) than a Placebo effect?

María Jiménez-Buedo[1]

Received: 8 May 2019 / Accepted: 25 January 2021 / Published online: 20 April 2021

Abstract
Reactivity, or the phenomenon by which subjects tend to modify their behavior in virtue of their being studied upon, is often cited as one of the most important difficulties involved in social scientific experiments, and yet, there is to date a persistent conceptual muddle when dealing with the many dimensions of reactivity. This paper offers a conceptual framework for reactivity that draws on an interventionist approach to causality. The framework allows us to offer an unambiguous definition of reactivity and distinguishes it from placebo effects. Further, it allows us to distinguish between benign and malignant forms of the phenomenon, depending on whether reactivity constitutes a danger to the validity of the causal inferences drawn from experimental data.

Keywords Dictator game · Experimental economics · Experimenter demand effects · Social science experiments · Hawthorne effect · Interventionism · Placebo effect · Reactivity · Validity

1 Introduction

The surge in social scientific experimentation of the last years has been in great part driven by the success of experimental and behavioral economics. It is natural then that the methodological discussions around new experimental practices in the social sciences have often been shaped by the debates that were taking place among practicing experimental economists.

In the first few decades after the emergence of new experimental practices in the social sciences, then, the question of reactivity, or the phenomenon that occurs when individuals alter their behavior because of their awareness of being studied, has not

✉ María Jiménez-Buedo
mjbuedo@fsof.uned.es

[1] Dpto. de Lógica, Historia y Filosofía de la Ciencia, UNED, UNED, Paseo de senda del rey 7, 28040 Madrid, Spain

been central to the discussions of methodologists or practitioners, partly because economists were not crucially concerned by it.

With their clear-cut methodological stance shaped most importantly by a tenacious control over the incentives faced by participants in the experimental setting, experimental economists may have initially felt that their experiments were shielded from the worries associated with subjects' reactivity that had long haunted their fellow social psychologist experimenters. More recently, experimental economists gradually moved in their study toward topics in which economic incentives no longer dominated the structure of a given game, but instead were intermingled with normative considerations (such as in the study of altruism, punishment, or social norms). Following this developments, a corresponding interest in the problem of reactivity has ensued among experimental economists.

In particular, the question of reactivity, under its multiple conceptual variants, has gained the attention of important experimentalists regarding the Dictator Game (DG)[1] and other similarly abstract designs aimed at measuring the normative inclinations of subjects. While the standard DG results, in which a number of "dictators" share their money with complete strangers has traditionally been interpreted widely as evidence of prosocial behavior, a number of important works that came out around the same time started disputing this interpretation, and instead suggested that the high level of donations observed was more likely indicative of the existence of artefacts: thus, for authors such as Bardsley (2008), Zizzo (2010), Dana et al (2007), and List (2007), the fact that a majority of DG subjects were willing to share a significant amount of their endowments with their fellow players was because the game was too transparently "about giving", and thus experimental subjects could easily guess what was expected of them and acted accordingly..

In this way, and according to critics, players in the DG are merely trying to perform the role of "good subjects" by adjusting their behavior to expectations, or, more specifically, adjusting to what they think it is expected of them as subjects (Bardsley 2008; Zizzo 2010). Alternatively, others have argued that relevant inferences from the DG and other similarly abstract games are still possible: both in the lab and in the field, subjects' behavior depends on other people's expectations and thus the DG provides a useful setting to study how subjects choose to adhere to the normative cues that the experimental setting provides (Levitt and List 2007, Jimenez-Buedo and Guala 2016).

Despite the shadow of the artifact over the DG, the game continues to be enacted in the growing number of social science experimental labs that have been set up in the last few years, coinciding with the extraordinary growth of experimental methods across the social scientific disciplines. There remains an open question regarding what can be inferred, if anything, from subjects' behavior in the standard DG or

[1] In the DG, the experimenter allocates some fixed quantity of money with player 1, the Dictator, who then has to decide how much, if any, he or she wants to share with player 2, the Respondent. The results of the standard DG show that roughly half of the Dictators depart from the earnings maximizing strategy and choose to give some money, the mean allocation being 20% of the initial endowment. Moreover, a consistent minority of dictators choose to split the sum in two similar sizes (Camerer 2003).

its variants. Can the DG results be used to explain phenomena outside the lab, and if so, which are those phenomena? Can we use the DG results to explain why people do things like give money to charities or is the behavior of DG players only meaningful (and relevant) *inside* the lab?

This paper argues that the debate about the validity of results of the DG and related games is stymied by the ambiguities that surround the concept of reactivity, as there are a number of unresolved conceptual issues regarding the phenomenon of reactivity. In this paper, we address two of these conceptual ambiguities.

First, there are a number of terms that are used to refer to what we here conceptualize as the phenomenon of reactivity, though they often are used without clarifying their definitions and, more importantly, they are often used interchangeably. In this way, Hawthorne effects, placebo effects, demand effects of experimentation, experimenter demand effects, methodological artifact, social desirability bias, are all terms that are often used in a loose way to invoke what we refer to as reactivity, or the phenomenon by which subjects in an experiment tend to modify their behavior in virtue of their awareness of being under study.

For example, and as we will see again in the next section, this is apparent in the debate around the validity of inferences from the DG, where an array of terms have been used often interchangeably to refer, in turn, to an array of ambiguously defined phenomena related to reactivity. In this way, and though there are many possible mechanisms for what we here call reactivity (such as the desire of subjects to comply with experimenter's expectations; their capacity to correctly guess the object of the experiment; the queasiness or apprehension of subjects to being evaluated; the fact that some subjects may try to deceive experimenters about their true motives for action; and the fact that experimenters or experimental designs may involuntarily give out cues about what behavior is expected of subjects), we here try to provide a unifying framework that subsumes the commonalities of these phenomena under the umbrella term of reactivity.

Second, and relatedly, there is the issue of whether the type of phenomena or mechanisms mentioned above invalidate an experiment's inferential import or whether instead, they only constitute a *potential* threat to the validity of inferences from an experiment. Again, because the definitions of terms such as Hawthorne effects, demand effects, and the like are often used without being standardized or operationalized, these terms are used interchangeably both to define the phenomenon associated with reactivity and to refer to the invalidation of an experiment's results due to the existence of reactive effects. This, again, creates confusion around the validity of experiments whenever we know or suspect that reactivity is at work in any given experiment. Here, we provide a framework in which we specify the conditions under which the existence of reactivity poses a threat to our capacity to draw causal conclusions from experiments.

In the pages that follow we provide a behavioral definition of reactivity. We offer an interventionist framework (Woodward 2003) that subsumes the phenomena associated with reactivity under a unifying conceptual scheme.. This framework allows us both to define unambiguously the notion of reactivity, and to analyse the challenges that reactivity can pose to causal inference in experiments with

humans.[2] To this avail, we introduce a distinction between malignant and benign forms of reactivity, in terms of the effects that reactivity can have on the validity of causal inferences drawn from experimental results. We argue that malignant forms of reactivity have the potential to render findings causally uninterpretable and we have reason to suspect that they do so whenever the effects of reactivity are idiosyncratic, i.e., whenever reactive effects cannot be assumed to be equal across the control and the treatment groups. Finally, our framework allows us to differentiate between reactivity and placebo effects.

Our paper also argues that clarifying this concept and the related set of phenomena that it describes therefore constitutes a valuable contribution to the debate about the limits of social science experimentation.

2 Reactivity again

In the early years of the experimental economics, when the focus was exclusively on the study of market institutions, experimental economists may have felt that their experiments were shielded from the worries associated with subjects' reactivity. This was due partly to the fact that experimental economics, born as a means to study the economic phenomena such as the clearing of markets, could adhere to a series of methodological principles, synthesized by Vernon Smith's precepts (Smith 1982),[3] and meant as a list of rules that provided sufficient conditions for the validity of experiments. Of these six principles, four of them were related to the need of adherence to strictly structured monetary incentives. Most importantly, the principle of *dominance* dictated that incentives had to *dominate* over any other subjective costs associated with participation in the experiment, thus creating a stark methodological barrier between the practices of economists and other more traditional experimental practices in psychology.

Gradually, the practices of experimental economists converged with those of behavioral economists (who themselves had a history of cross-collaboration with psychologists) and this convergence crystallized in a methodological synthesis in which there was a clear relaxation of some of the Smithian precepts. Yet, there was still the perception that economists and psychologists differed systematically in their methodological practices, as summarized in the classic Hertwig and Ortmann piece

[2] It is perhaps opportune to underline once more that the framework for reactivity we provide, in which it is used as an umbrella term, does not intend to distinguish among different mechanisms of reactivity. It instead unifies the phenomenon in order to explore the problems that it can create to causal identification and experimental validity.

[3] Vernon Smith's precepts were the following: the proscription of deception, the principle of parallelism, or the idea of "similarity" between the lab setting and the target phenomena, and finally, a series of requirements regarding the structure of the incentives faced by subjects. These included: (i)nonsatiation (where the medium of payment should not "satiate" participants, in the way, more money does typically not satiate); (ii) saliency (where the reward must increase or decrease according to the way in which an outcome is consider good or bad, or correct or incorrect); (iii) dominance (where the rewards must dominate any subjective costs associated with participation in the experiment), and (iv)privacy (in that each subject in an experiment receives info only about her own payoffs).

(2001). Following Hertwig and Ortmann, these practices (the proscription of deception, the use of well-defined scripts, and the repetition of tasks), together with the use of monetary incentives, were defining features of the experiments in economics, as compared to those of psychologists. None of these practices were in themselves warrants against reactivity but they may have, collectively, during some years, given a sense of protection against the perils of reactivity to a profession that was gradually and increasingly adopting experimental practices.

As experiments became common within the discipline of economics, experimenters in economics broadened the array of topics that they dealt with. Gradually their topics included, prominently, questions regarding pro-social behavior, but in these games, by construction, monetary incentives needed to be weighed against other (pro-social) considerations: they could no longer completely *dominate* the incentives of the players (Jimenez-Buedo 2015). Against this background, and as we already pointed out in the introduction, the success of games such as the DG and the ensuing debate over the correct interpretation of its results eventually brought the question of reactivity to the fore of the methodological discussion among economists.

Initially, the critics of standard interpretations of the Dictator Game results resorted to standard terminology used in more traditionally experimental disciplines, such as psychology. For example, as already mentioned above, Bardsley (2008) resorted to the concept of Hawthorne effects in his criticism of altruistic interpretations of the DG results. The term of Hawthorne effects, with origins in industrial organizational studies, is normally used to refer to the fact that subjects may try to "overperform" when they are being observed.[4] Because its definition is not standardized, it is also often used to refer to the subject's sensivity to being observed and sometimes also to refer to the behavioral changes that are considered to be a direct response to the experimenter's scrutiny.

Among the DG critics, Zizzo (2010) provides his own conceptual approach to the issue, and coined what is now the standard terminology in economics. Zizzo defined experimenter demand effects (2010) as the changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior. According to Zizzo, experimenter demand effects can be either purely cognitive (when an experimental participant tries to figure out what she is expected to do as an experimental subject), or they can also have an additional social layer, when that elucidation is additionally shaped by a sense of social adequateness.

Moreover, Zizzo's conceptual scheme also provided an account of the way in which experimenter demand effects could affect the validity of experiments. According to his framework experimenter demand effects are a problem for the validity of experiments whenever experimental participants can correctly guess the true experimental objectives.

[4] The origin of the term comes from the Hawthorne Works, in Illinois, a factory in which, in the context of a series of studies on productivity, a group of assembly employees seemed to paradoxically increase their productivity as researchers dimmed the lights. This puzzling result was interpreted as the result of the perception of workers of being under study (Adair 1984).

The term *experimenter demand effects* has been very extremely influential among experimental economists, and due to the influence of economics in the new wave of social science experimentalism, it is already permeating the language of experimentalists in other social sciences, such as sociology and political science, thus constituting the new conceptual standard. The term *experimenter demand effects* constitutes in itself a sort of terminological synthesis with respect to pre-existing terms in social psychology, by merging two classic terms: experimenter effects, and demand effects of experimentation. These other two terms constituted two important tenets in the lingo that originated in social psychology in the 1960s and 1970s and that has conformed, for years, the vocabulary of social scientific experimentalists: the synthesis would come from merging together, in one term, Orne's demand characteristics of experimentation and Rosenthal's experimenter (expectancy) effects. In the case of the former, Orne (1962, 1969) studied, both theoretically and empirically, how experimental subjects actively contribute to complete and construe the experimental task by enquiring and hypothesizing what is expected of them as experimental subjects. For Orne, this is an inherent feature of social scientific experimentation, since experimental instructions are necessarily incomplete: the experiment is itself a social situation that exerts implicit demands on the social actors involved in it. These implicit demands are worthy of study by social psychologists (thus, Orne's and other's project of a Social Psychology of Experimentation). More practically, Orne also considered that these demands need to be analyzed by experimenters because they have the potential to interfere with the (more explicit) experimental task that is the experimentalist's primary object of research. Rosenthal's experimenter expectancy effects (1968), in turn, refer to the set of cues regarding the experiment's objectives or hypotheses that experimenters can inadvertently send to participants, and that can end up affecting the experiment's results. In this way, and by focusing on experimenter demand effects, Zizzo merges both of these traditions in how he conceptualizes these effects: these are changes in the behavior of experimental subjecs due to (experimenter) cues about what constitutes appropriate behavior ("demanded" from them).

Zizzo classifies demand effects on the basis of whether subjects correctly or incorrectly guess the true goal of the experiment. Thus, depending on the coincidence between what the subjects believe about the experiment and what the experiment really is meant to test, we have three possible cases:

1. Uncorrelated expected and true objectives
2. Negatively correlated expected and true objectives
3. Positively correlated expected and true objectives

Zizzo argues that only the third case is truly problematic: demand effects in this case act as a confound, preventing the researcher from distinguishing the causal role of the treatment from that of the demand. This is, according to him, the case of the standard Dictator Game: the experimenter's demand is correlated with the true purpose of the experiment, because subjects can easily guess that the experiment is about "giving." Zizzo's terminological effort is commendable,

among other things, for trying to offer an account of the conditions under which experimenter demand effects affect the validity of experiments. But Zizzo's specification remains unsatisfactory for the reasons discussed below.

A look at some standard practices in the more orthodox practices of experimental economics suffices in order to see why Zizzo's diagnosis regarding the effects of experimental demand effects on validity lacks generality: monetary incentives (especially when or if they are dominant) are often used, precisely, to align the motivation of experimental subject with the (true) objectives of experimenters in a given game. In other words, they are used to signal to participants what the real objectives of a given experiment are. This is the case, for example, in those instances in which experimenters create an environment where income maximization is expected and demanded from participants. The coincidence between the true experimental objectives and those guessed by participants is in these cases, rather than a problem, a precondition for success in the experiment. This is a weakness in Zizzo's diagnosis regarding the relation between reactivity and the validity of experiments.

The interventionist account that we introduce next avoids this problem by bypassing any reference to the "experiment's true objectives ", a notion that can be vague and hard to operationalize. Yet, our account still provides a way to distinguish between situations in which reactivity is not problematic for experimental validity versus situations in which it potentially poses a threat. As we mentioned in the introduction, to properly discern between these two situations is important terminologically, as this is one of the ambiguities that hinders discussions on reactivity by producing misunderstandings: most of the terms that we use to refer to the general phenomenon of reacitivity (such as experimenter effects, demand effects, placebo effects, Hawthorne effects, or methodological artifacts) are often used without distinguishing between two different aspects of the phenomenon: these terms are used to refer *both* to the *mechanisms* that have the potential to bias an experiment and to the *biases* that can (or not) result from these mechanisms.

As we have already mentioned, some of the terms that are normally linked to reactivity-related phenomena have, in some contexts, some more specific meanings. This is the case, for example, for the term Hawthorne effects, which in some contexts can refer to the fact that experimental participants often feel motivated to display their best performance at a given task (and in this sense, better than they would under normal conditions), as a result of their being under study. Yet, in other contexts (as was the case in the DG debate), the term is also used in a different sense, to refer to the participants' motivation to adapt their behavior to whatever they think the experimenter expects of them. While these two different types of attitudes can coincide in some contexts (e.g., whenever experimenters expect participants to perform at their "best" and subjects anticipate it), there are scenarios in which these two types of participant attitudes would lead to diverging behavioral responses.[5] For this reason, using the same term to refer to both phenomena can lead to confusion.

[5] Note, as an example of how the two phenomena may differ in a concrete example: In the original Hawthorne Works study, employees responded by overperforming as lights became dimmer, though it is unlikely that they would have thought that experimenters expected them to do so.

Here we defend an approach that unifies all reactivity-related phenomena under the same label, by focusing on the common aspects of the different mechanisms that can lead to reactivity. This does not preclude that further studies focus on more specific mechanisms, but rather, we contend that in an area where terminological ambiguity abounds, providing first a unifying framework is a useful first step.

3 Interventionism and social scientific experiments

In this section we characterize reactivity and the challenges that it poses to causal inference by using an interventionist or manipulationist account of causation (Woodward 2003, Spirtes, Glymour and Scheines 2000,[1993]). For this, we will first describe the basic tenets of causal interventionism to then characterize a common type of behavioral experiment using an interventionist framework. We then analyze the possible meanings of reactivity through an interventionist lens.

An interventionist conception of causation conceives causal relationships as relationships that describe what will happen to some variables (effects, or dependent variables) when we manipulate or intervene on others (causes, or independent variables). For an interventionist to say that a relationship is causal is thus to say that it is exploitable for purposes of manipulation and control in a way that merely correlational relationships are not. The choice for this framework given our present problem (i.e., reactivity and how it affects causal inference from experimental data) seems natural for three reasons:

First, the interventionist notion of cause is often justified, precisely, as one that is especially fitting to the logic of the controlled experiment, which in turn is regarded as a method privileged in its capacity to allow for the testing of causal claims (pp. 22–23 Woodward 2003). In fact, interventionism can also be interpreted as a methodology to find out about causes, rather than as an approach committed to any particular ontology of causation (see Woodward 2015). Understood as a methodology, interventionism associates causal claims with the outcomes of hypothetical *experiments* in which the value of the variable representing the putative effect is set by means of intervening (only) on the putative cause.

Second, interventionism as conceived by Woodward has been especially concerned with the identification and clarification of ambiguous causal claims as they come up in (often social) scientific contexts, such as the assertion that "being female causes one to be discriminated against in hiring/salary" (p. 115, Woodward 2003). Woodward has tried to clarify such claims by linking them to potential or actual experimental manipulations. As we will show, the ambiguity in some of the assertions involving the phenomenon of reactivity comes, precisely, from a lack of clarity regarding what types of manipulations are attainable in different experimental settings involving humans.

Third, although Woodward has dealt with psychological and social science experiments that study social preferences (2007, 2008), the question of reactivity has not been systematically analysed under an interventionist framework: though Woodward has studied some well-known economic experiments such as the Ultimatum and the DG, his discussions have dealt with the robustness and external validity of their findings, but he has not, to date, specifically dealt with the phenomenon of reactivity and the question of how it can affect the causal claims we can validly infer from

these games. The present paper thus contributes both to the literature on interventionism in social scientific experimentation and more broadly to the methodological and philosophical debates around experimental social science.

According to Woodward's well-known manipulationist definition of cause:

(M) X causes Y iff (1) it is possible to intervene on X and (2) under some such possible intervention on X, changes in the value of X are associated with changes in the value of Y. Interventions must in turn fulfill the following conditions (see Fig. 1):

IN-i The intervention I completely disrupts the causal relationship between X and its previous causes. The value of X is set entirely by I.
IN-ii The intervention I should not itself be produced by any process that affects Y via a route that does not go through X.
IN-iii The intervention I leaves the values taken by any causes of X except those that are on the path from I to X to Y unchanged.
IN-iv The intervention I must not directly cause Y via a route that does not go through X.

In more recent work Woodward (2007) has relaxed condition IN-i, which defines hard or arrow-breaking interventions in order to accommodate processes in which the value of X does not come entirely under the control of the intervention. This happens when there are other endogenous causal influences on X that cannot be broken by the intervention. In those cases. IN-i can be relaxed to IN-i', where *the intervention supplies an appropriately exogenous and uncorrelated source of variation to the variable X* intervened on, rather than a complete disruption or breaking of all other causal influences on X. Thus, in soft interventions thus defined, the variation supplied by the intervention I should not be correlated with other causes of X or with causes of Y besides those that are on the route from I to X to Y.

The relaxation of this condition is crucial to accommodate experiments in many areas in which proper surgical interventions are not possible. In the case of the behavioral sciences, the impossibility is often determined by the fact that some form

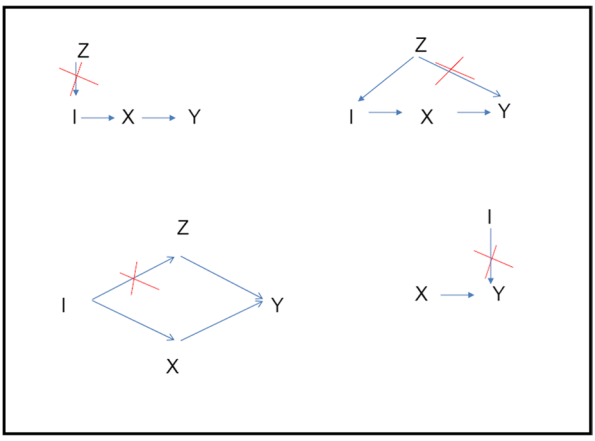

Fig. 1 Conditions IN-i to IN-iv for an ideal intervention (left to right, top to bottom)

of mental causation is involved: as it has been argued by Campbell (2007), condition IN-i would entail that whenever we want to intervene on the mental state of an agent, we must ensure the removal of all the other causes of that agent's mental state (thus suspending the rational autonomy of the individual).

Now that the main elements of an interventionist framework are laid out, we can use it to represent some economics experiments. In particular, we want to focus on the type of experiments that have sparked some of the recent discussions about reactivity in experimental economics. For this reason, we will use the DG as an example, as it is a well-known game with a very simple structure facilitating exposition, and has the additional advantage of having been extensively discussed by leading experimentalists in regard to reactivity-related issues.

By introducing modifications to the basic structure of the game, The DG design has been used to test different types of hypotheses. Here we focus on a well-established use of the DG design: the testing of subjects' sensitivity to the manipulation of the normative framework applicable to the experimental situation (Guala and Mittone, 2010). Typically, in this kind of experimental exercise the basic DG is played as a control against a modified DG that constitutes the treatment, where the modification consists of the introduction of a normative-relevant cue. For example, in a well-known example, subjects in the treatment group play the DG in a room in which a picture of a pair of eyes is set, in order to bring to the subjects' imagination the possibility of someone observing their actions (Haley and Fessler (2005)). Other well-known modifications of the DG include introducing a modification in the identity of the Recipient (from an anonymous player to a well-known NGO, for example), or introducing an element of merit in deciding who, among two given players, gets to be the Dictator.

To be sure, both the standard DG (acting as a control or baseline) and the modified DG (acting as the treatment of interest) expose experimental subjects to an "unusual" normative setting, but the assumption is that by further modifying the normative environment, we can test whether an additional normative cue further affects the subject's willingness to donate. The difference in the mean allocation between the two games is then interpreted as reflecting the impact of the introduction of the experimental manipulation in the modified DG: in terms of the causal hypothesis being tested, the difference in the mean allocation (from Dictators to Recipients) in the two experimental settings is seen as being caused by the introduction of the normative cue.

We can thus conceptualize this experiment, in more formal terms, as one based on a double intervention, where we must compare the results of each intervention to draw a conclusion about the causal impact of our putative cause on the putative effect (or the impact of the independent variable on the dependent variable). For this, we compare a control group playing the standard DG (X_0), with a treatment group exposed to the introduction of a DG that includes an additional normative cue (X_1). The causal impact of the normative change in the environment (X_1-X_0) is thus measured by the difference in the mean allocation (Y_1-Y_0) See Fig. 2 below.[6]

[6] The representacional convention (I = on and off) is borrowed from Eberhardt and Scheines (2007).

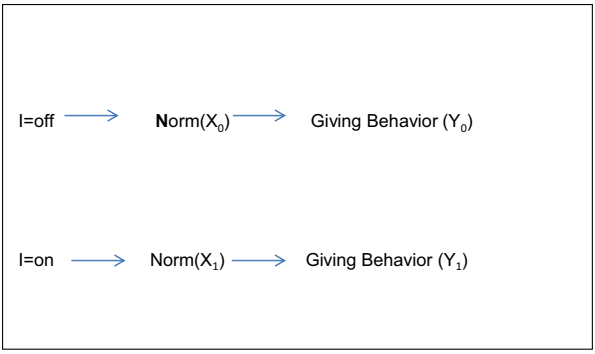

Fig. 2 The Dictator Game from an Interventionist Perspective

We are now in a position to offer a suitable conceptualization of reactivity from an interventionist perspective. Recall that reactivity does not need to be restricted to experiments, as it is usually understood as the change in the subject's behavior that results from his or her awareness of being studied, where this is also applicable to observational studies. In an observational environment, the change in behavior will come as a result of a subject´s awareness of being studied, or as a result of the operation of whatever measurement device is used. In the case of experimental studies there is an added layer of complexity, since by its own nature, the experiment provides the subject with a stimulus that is often supposed and expected to cause a behavioral change in participants (by exposing them to the treatment, or putative cause). Thus, when specifically applied to experiments, most definitions of reactivity-related phenomena can be seen as somewhat elliptical: reactivity is the change in the subject's behavior as a direct result of her being studied, *rather than* as a result of the operation of our variable of interest, although the second part of the sentence is often not explicitly mentioned.

In terms of the categories deployed in an interventionist scheme, reactivity can thus be defined as a byproduct of an experimental intervention due to the subject's awareness of taking part in that intervention. This byproduct takes place outside the causal path that goes from the independent variable or putative cause to the dependent variable or putative effect: we intervene on X (the putative cause) in order to assess its effect on Y (some aspect of the subject's behavior), but by *intervening* experimentally, we also affect the subject's behavior via some other route that does not go through X (i.e., the subject´s behavior gets altered because of his or her awareness of being under study).

Figures 3 and 4 represent cases of reactivity associated with experiments with settings akin to that of the DG: reactivity occurs when an intervention produces a change in the subject's behavior through a route different from the one that goes from the putative cause to the putative effect (from I to X to Y).

Let us illustrate this definition with our DG example, where an intervention introduces a normative cue in the environment in order to test for its causal effect on the subject's "giving behavior". Reactivity would occur if the intervention also results in inducing in the participant, for example, a sense of apprehension (such as a sensation of queasiness over feeling observed or studied upon) and if, in turn, the participant reacts to this apprehension by modifying his behavior (such as, for example, sitting up straight in his chair as a response to the feeling observed). Note that the

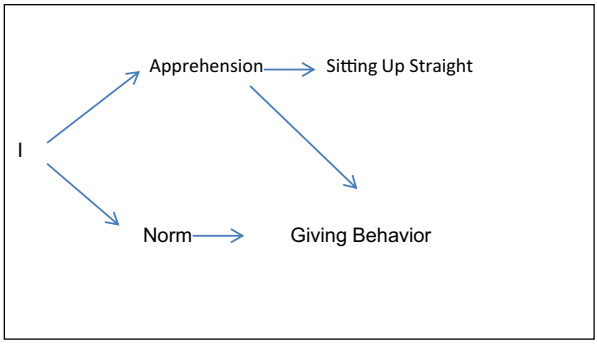

Fig. 3 An example of Reactivity

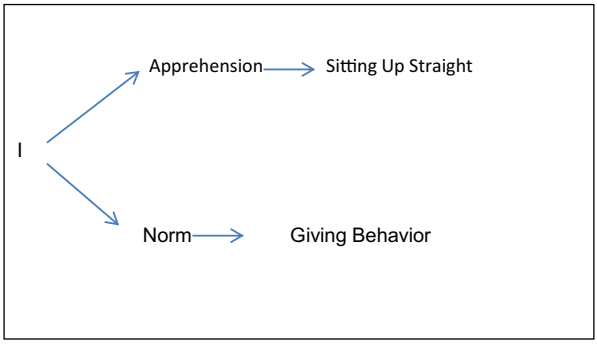

Fig. 4 Another example of Reactivity

apprehension is not attributable to the introduction of the normative cue per se, but to some other aspect imbued in the experimental setting, such as the fact of being under observation (see Figs. 3 and 4 above). It should be noted that apprehension to evaluation is only one of the many potential triggers of reactivity, where other common, well-known manifestations or mechanisms include the subjects' reactions to the perceived authority of the experimenter, the participant's zeal for being "a good subject" (or the opposite uncooperative desire to "boycott" an experiment), or the pervasive and understandable participants' active search for cues and second guesses about what *the experiment is really about* (Jimenez-Buedo and Guala 2016).

By conceptualizing the phenomenon of reactivity in this way, we can better see what distinguishes reactivity in an experimental context from the more encompassing, general phenomenon of reactivity in observational research. Reactivity occurs when by studying subjects, we modify their behavior. However, in an experimental context there is always an intended intervention on the subjects' environment, often purposefully directed at behavioral change. Reactivity is thus the uncontrolled, unintended effect on the subjects' behavior that results as a byproduct of the intervention put in place to test for the causal effects of the experimental treatment. As we will see in the next section, our interventionist framework allows us, precisely, to discern

when and why the intervention's behavioral byproduct poses risks to our capacity to draw causal inferences from the experimental data.

4 Benign and malignant forms of reactivity

Now that we have defined reactive behavior within an interventionist framework, we can distinguish between two types of reactivity, depending on whether the type of reactive behavior violates or complies with the conditions for an ideal intervention.

Benign reactivity occurs when the intervention's impact on the subject's behavior does not affect the output variable of interest in the experiment. It is thus *benign*, in the sense that it does not pose in itself any problems to the causal inferential process as conceived by interventionism. Figure 4 shows an example of benign reactivity: intervening to set the value of the putative cause triggers an additional behavioral effect (sitting up differently than we normally would). This effect, however, operates outside of the causal path going from X to Y, and does not affect Y in any way.

By not violating any of the conditions of an ideal intervention, benign reactivity does not pose any particular challenges to causal inference. In our DG example, benign reactivity would mean that the apprehension that DG players can experience causes them to sit differently in their chairs (or makes them more prone to smiling, or causes their heart to beat faster) but to retain its benign character that same apprehension *cannot* affect the players' "giving behavior".

We can define malignant reactivity, in contrast, as occurring when the experimental manipulation not only changes the value of the putative effect Y by setting in motion the putative cause X, but additionally, it gives rise to an additional causal path that also affects the output variable of interest Y. This violates condition IN-iii above, so manipulations in which malignant reactivity occurs do not constitute ideal interventions in the Woodwardian sense.

Figure 3 represents graphically a case of malignant reactivity: the intervention sets in motion some reactive mechanism in Dictators (such as apprehension) and this apprehension affects, in turn, their willingness to donate to Recipients. In this case, the Dictator's donating behavior is influenced both by the manipulation of the normative framework and by the participants' apprehension toward the experimental evaluation of their behavior. Malignant reactivity thus constitutes an obstacle to causal inference through the violation of the IN-iii condition: if the level of donations we observe is suspected to be due not only to our introduction of a normative cue (the putative cause) but also influenced by some concomitant factor (in this case evaluation apprehension), then the effect that we observe on donations when we intervene on the normative cue cannot be attributed solely to it.

Note that the introduction of the distinction between malignant and benign forms of reactivity solves an extant ambiguity in the way that the relevant literature treats the relation between reactivity and experimental validity: the many terms that are employed to refer to reactivity-related phenomena are normally used to designate both the phenomenon itself *and* its potential for undermining the validity of experimental inferences. In this way, it is often the case that terms such as Hawthorne effects, are used ambiguously to refer both to the phenomenon by which a subject, for example, may be motivated to

perform his or her best in an experimental context, and to refer to the experimental artifact that a particular reactive behavior may cause in a particular experiment. The problem with this ambiguity is that if it goes unnoticed it implicitly amounts to assuming that any reactivity-related phenomenon *ipso facto* invalidates any experimental inference that we wish to make. Yet, the two need not go together, as we might well be in situations in which, for example, we want, as experimenters, to motivate participants to perform at their best level, having no reason to think that their doing so poses a problem to the validity of our inferences from the experiment.

Because we also know that some form of reactivity or another is always present in any social scientific experiment, the implicit automatic connection between reactivity and artifact is likely to play no small role in the thinking of those that see social scientific experimentation as an enterprise doomed to fail. Yet, most social scientists and commentators tend to think, more plausibly, that reactivity does not irremediably lead to the invalidation of an experiment, yet the systematic discussion on the conditions under which it would are often absent.

In this regard, Zizzo's more ambitious conceptual project is careful: in his framework, experimenter demand effects are not in themselves a problem but have the potential to create one whenever experimental subjects can correctly guess the objectives of the experiment, yet, as Jimenez-Buedo and Guala (2016) have argued, this approach neglects that often many economic experiments successfully align the incentives of subjects and experimenters through monetary rewards that are meant, precisely, to inform experimental subjects what exactly is sought of them, or in other words, what the objective of the experiment really is. Thus, and although Zizzo's identification of this condition seems to fit the DG case nicely, it does not constitute the best grounds for a general elucidation of these conditions.

Our definition of reactivity and our distinction between benign and malignant forms of reactivity solves this problem: reactivity can but does not necessarily cause problems for causal inference. In its benign form, reactivity does not in itself pose difficulties in terms of the causal inferences that we can draw from experiments. In contrast, malignant reactivity constitutes an obstacle to the inference of causality from experimental data.

5 Is malignant reactivity lethal to causal inference? Placebo effects versus reactivity

The previous section ends on a somber note regarding the damage that malignant reactivity can do to experimental exercises aimed at inferring the causal impact of a given variable through controlled interventions. Yet, the reader may immediately consider the parallels between malignant forms of reactivity and what routinely occurs in Randomized Controlled Trials when placebo effects are present (i.e., when expectations about treatment have an effect on the recovery of patients). After all, the interventions that normally take place in RCTs often include, via placebo effects, a violation of condition IN-iii: the placebo effect created by exposure to *any* treatment (active or placebo) can improve our mood or expectations in ways that in turn impact our health. Yet, as we know, the introduction of control groups routinely solves whatever problems this may create for causal inferential purposes.

In fact, Woodward (2008) has discussed how an interventionist account can analytically deal with the presence of placebo effects in drug-testing RCTs. He has done so in the context of his response to Cartwright's criticism of the interventionist's assumption of modularity. Woodward argues that even though, as Cartwright rightly points out, placebo effects make surgical interventions impossible, interventionism can account for the strategies employed for inferring causality despite the impossibility (2008, p. 212)). In the presence of the placebo effects an interveniorist approach provides the rationale for the introduction of a control group that receives a placebo (a drug that resembles the treatment in all but its active ingredient). The aim of this placebo control group is to provide a base-line that allows us to measure the net causal effect of the drug we are testing by means of comparing the output result in the control trial and drug trials. The difference between the two trials is thus assumed to be an accurate representation of what *would have happened* if the drug had been administered in the absense of a placebo effect. In an interventionist account this subtraction or net effect represents or stands for the results of a counterfactual trial in which a surgical, ideal intervention would be possible. If the solution is readily available in the case of placebos, can we not use it to deal with the case of malignant reactivity in social scientific experiments?

In Fig. 5 we can see how the structure of the problem is formally similar in both the DG in the presence of reactivity and in an RCT with placebo effects. In both cases we see how malignant reactivity is present. However, there is a crucial difference between both situations: whereas in the case of RCTs the assumption of placebo effects that are equivalent across treatments seems generally valid (or at least valid for all those experiments in which the treatment can be administered in ways where blinding is effective,[7] such as in the intake of pills), it seems much harder to satisfy in the *case of treatments* involving some form of mental causation.[8]

The reason is that in the case of social scientific experiments, a given treatment (or placebo) needs to be embedded in an experimental script, to which subjects then react. In some ways the experimental script carries the variable of interest like a pill may carry (or not) an active treatment: the variable of interest (say, a normative cue) is embedded in a given script like an active principle is embedded in a pill. Yet, this "carrying" also differs in important ways: in the case of experimental treatments involving mental causation, the script that "carries" a given treatment also *embodies* it, in a way in which the script and the treatment in which it is embedded become an inseparable bundle to which the subject reacts. For this reason, whatever reactive behavior occurs, it is likely to be the joint product of all the experiment's elements in conjunction and this, in turn, implies that each script has the potential to give rise to its own unique, idiosyncratic reactivity: even if

[7] For an analysis on the relevance of blinding see Teira and Reiss (2013) Teira (2019).

[8] To be sure, placebo effects *also*, and rather obviously, involve mental causation, but on this point we are contrasting the *treatments* that are being administered, not the secondary effects (both of which -placebo and reactivity alike- involve mental causation). In the case of the social sciences treatments, they will almost always involve some form of mental causation, in contrast with the case of the medical placebo if the treatment is administered through the intake of a pill.

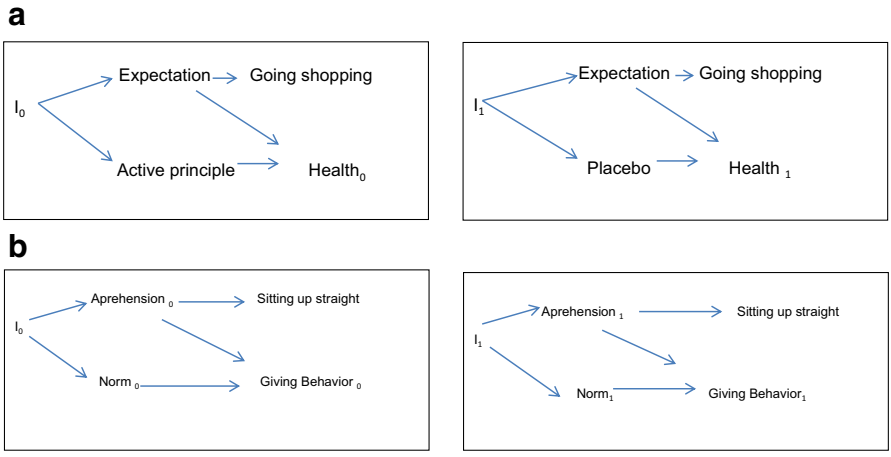

Fig. 5 **a** RCTs and the placebo effect. **b** Reactivity in social scientific experiments

the treatment and control protocols differ in only one element (i.e., the presence or absence of our intended independent variable of interest), we cannot rule out that this differential element is enough to alter the participants' perception of the whole experimental experience. This means that even the part of the script that remains the same across treatments can be perceived differently (as part of a different whole) by the experimental subjects.

When we add an active principle to a pill in the control group, the active principle alone can explain the difference between the responses in the treatment and control groups. In contrast, when we add (for example) an additional normative cue to an experimental script, the difference between the respones in the treatment and control group is the result of the interaction of the normative cue with the script. Put in other words, the inclusion of an element whose causal impact we want to test (e.g., a normative cue) has the potential to modify the effect of the *same* base script across the experimental groups, since the normative cue and the script that embeds it will be received inseparably by the experimental subjects. The same script used on its own (in the control group), and used in conjunction with the treatment (in the treatment group), might be received differently. This stands in contrast with the case of an RCT testing the efficacy of an active ingredient: once we assume that blinding across treatments is effective, we can safely assume that the excipient in the pill ha the same (placebo) effect across the treatment and the control groups.

In social scientific experiments, when a design tries to isolate the causal effect of a treatment embedded in a script, we must however at least conceive of the possibility (in cases where we suspect that there is malignant reactivity) that the differences in behavior across groups may be due not only to the treatment itself (understood here again as the variable of interest) but also, that this difference across treatments may be also due to the differences (across treatments) in the reactive behavior. This means that even if we introduce a minute change in the treatment group (minute with respect to the control group), we may also be modifying differentially across

treatments, things like the participant´s eargerness to cooperate with what she thinks is the experiment's objective, or her apprehension to the experimenter's evaluation.

The reason for this lies in the holistic nature of meaning in social interactions: because any minute difference in a script has the potential to alter the meaning of a social interaction, a small difference in a script can transform the subjects' interpretation of the experiment and thus can change the reactive behavior associated with it.

This has an important implication for social scientific experiments aiming at testing causal hypotheses through the comparison of control and treatment groups: if we cannot assume generally that these two interventions give rise to the same type of reactivity, then we cannot assume generally that a standard control group will suffice in order to correctly identify and isolate the causal impact of treatments net of reactivity. This will be the case even i (as it is often the case), the control and the treatment differ in only one minute element, for that minute element has the potential to change the interpretation of the whole experiment and to induce different types of reactivity in both the control and the treatment groups. As we have shown, this aspect of social scientific experimentation can be well represented and conceptualized through an interventionist framework.

This paper thus clarifies the phenomenon of reactivity by subsuming it under this well-known framework. An interventionist framework allows us to provide a behavioral definition of the phenomenon of reactivity, subsuming its different mechanisms under a general scheme. It allows us, further, to distinguish between benign and malignant forms of reactivity, by differentiating between situations in which reactivity affects the variable of interest, from those in which the reactive behavior is orthogonal to the variable of interest.

In this section we have also seen how an interventionist framework can allow us to differentiate between situations in which malignant reactivity can be remedied with a control group (as it is routinely the case in RCTs dealing with placebo effects), from those situations in which malignant reactivity may be "resistant" to the standard procedure of contrasting the treatment and control groups. The latter can happen whenever reactivity may be idyosincratic, meaning that it is unique to the particular script enacted in each experiment. If reactivity is of this type, it cannot be subtracted away by comparing the treatment and the control group, even if the treatment and control differ in only one element. Summing up, an interventionist framework thus allows us to show that experimental reactivity can pose a threat to the inferential import of experiments. According to this framework this will happen in cases in which this reactivity is both malignant and idiosyncratic.

An interventionist framework thus provides a clear account of cases in which reactivity is present, but benign to the validity of an experiment, and it further provides a clear account of situations in which, in contrast, reactivity poses a threat to validity even if we have a (placebo) control group.[9] This contrasts with previous

[9] It should be noted that malignant idiosyncratic reactivity threatens not only external validity, but also, internal validity. Regarding external validity, the existence of reactive effects that can not be substracted away via a control group, undoubtedly poses problems to the extrapolation of results from the lab to non experimental conditions. The problem, however, is also one of external validity: proper causal identification through isolation is not possible in the presence of malignant idiosyncratic reactivity, and thus, internal validity cannot be attained. I thank reviewers for pressing me on this point.

analysis of some aspects of the phenomena, and especially, with Zizzo's account of experimenter demand effects, in which they are supposedly a threat to validity in cases in which experimental subjects can correctly identify the true objectives of the experiment.

Let us illustrate this analysis with our example contrasting the use of placebo in a properly blinded RCT with the case of a DG in which we assume malignant idiosyncratic reactivity (examples also depicted in Fig. 5):

If in an RCT set up to test the effectiveness of a new drug a given participant's mood is improved merely by taking part in the study (i.e., if he or she is subject to a placebo effect), then we can safely assume that this improvement in mood will be equivalent across the treatment and control groups, in so far as blinding of the treatment is effective.

In contrast, consider the case of a standard DG used as a control and a modified DG used as the treatment of interest. If a participant is feeling apprehensive regarding the scrutiny of her behavior in a standard DG, this apprehension will be linked to her interpretation of the experiment's meaning, which in turn will be determined jointly by her overall experience as a participant, i.e., by all the elements consisting of the experimental setting. In the standard DG subjects might feel queasiness regarding the fact that the standard DG is a "mysterious", or an unusual game, where it is not totally clear what sort of behavior is expected of them. If we add an additional stimulus to the game in a modified DG (such as, for example, revealing the identity of the Recipient as being a charitable organization) we may, as experimenters, be using this stimulus as the carrier of a normative cue, the effects of which we want to test. However, the stimulus will also be the likely carrier of its own particular form of reactivity, one that has the potential to differ systematically from the type of reactivity associated with a standard DG. A modified DG can perhaps provide clearer signals to participants about the normative expectations at play, thus turning the environment into a more familiar one. At the same time, however, the range of phenomena linked to reactivity, (i.e., the behavioral response that is due to elements other than the intended treatment) is also likely to differ from that of a standard DG, and might, for example, have more to do with uncontrolled expectations regarding how to appear as a good subject.

In other words, to the extent that any two treatments involving social interactions are different (e.g., the baseline and the treatment of interest) we can expect (or at least consider the possibility) that their associated reactivity can be, in principle, unique and intrinsic to each treatment. The methodological consequence of this is clear, and applies as well to our DG example: the difference in the donation levels across treatments (the output variable of interest net of the baseline or control) can *not* thus be automatically assumed to be an accurate representation of what would have happened if the treatment of interest had been administered in the absense of reactivity.

To sum up, a variation in the script needed to modify a standard DG in order to carry a treatment (as, e.g., when introducing a normative cue in a modified DG) is likely to carry with itself a new bundle of reactive phenomena. If this reactivity is of the malignant sort, i.e., if it carries behavioral effects onto our output variable of interest, and, if we think it is idyosyncratic (i.e., if we think it depends on

the particular script we are enacting), then we may not have any obvious means to know what would be the effect of our treatment variable, net of reactivity.

And yet, in the case of the DG, a significant difference in means between a standard DG (baseline) and a modified DG (treatment) is routinely presented in the relevant literature as proof of the effect on donations of whatever modification in the game. It should be noted that this implies that this difference in means can be interpreted as representing the effect of the introduction of the normative cue (the treatment) on donation levels, net of reactivity. However, as we have shown, this operation rests on endorsing at least one of the assumptions below:

a. There is no reactivity involved either in the standard DG or on its modified version.
b. Whatever reactivity there is, it is of the benign sort for both the standard DG and its modified version.
c. If there is malignant reactivity on the DG or its modified version, this malignant reactivity is behaviorally equivalent in its impact on the variable output of interest (the level of donations), i.e., it is not idysincratic to the treatment.

While any of the above assumptions can in principle be true for any given experiment, they cannot be assumed to hold generally across all social scientific settings, especially in cases in which we have reason to think that some forms of reactivity are likely, as in the case of the DG and related games. Our framework shows why it is necessary to justify or discuss each of these assumptions in every instance, and for each intervention, when presenting social scientific experimental results.

Note that the case in which reactivity is *both* malignant *and* idiosyncratic is the truly challenging one, for what we call here malignant reactivity can otherwise be routinely treated through the use of control groups, as it normally is. Our goal here is to provide an account of reactivity that can clarify why these situations can happen (and why they cannot be solved by the standard practice of having control groups). Our aim here is theoretical and conceptual rather than stritctly practical, meaning that we try to provide the definitions and distinctions that can be of help to further research aiming at systematically articulating what concrete experimental settings tend to bring about these problems. Though our aim is not here to provide a guide that identifies the concrete conditions under which reactivity can be either malignant or malignant and idosyncratic, we can hypothesize that there are a number of experimental situations where we can suspect that we are in this predicament. In particular, the DG can provide some cues regarding some of the scenarios that can make reactivity of the malignant idyosincratic kind more likely to emerge.

The DG provides an example of a setting where we have a game that, having very little structure, produces very different results depending on the introduction of different cues or variations in the context. Put differently, the interpretation of the DG's "meaning" seems to depend on minute context variation. We can tentatively hypothesize that scenarios where results are very "sensitive" to slight changes in the experimental script might also be candidates for being scenarios where slight changes in the script can bring about strong changes in the part of the behavior that is properly

"reactive". In these cases, we might suspect that the reactive behavior might not be the same across the treatment and the control groups, provided that we think that the sensitivity of the design affects, not only the behavior in the treatment's causal path, but also, the part of the behavior that is properly "reactive".

As discussed in the introduction, the DG is a game in which, by construction, monetary incentives in the game do not *dominate* behavior (needless to say, if they did, the DG results would be incredibly boring, with zero donations across the board, irrespective of the particular designs). It seems to us that the DG exemplifies one of the obvious costs of abandoning *dominance* as a methodological precept: when economic incentives do not dominate the game, there is room for other considerations, including "reactive" ones, to affect the behavior of the participants in an experiment. But abandoning the principle of *dominance* is necessary if economists are interested in studying social behavior that relates to normative or ethical motivations, for the study of these through monetary incentives is done, precisely, by weighing monetary incentives against these other social (e.g., purely normative) considerations. In this sense, economists, once they have abandoned dominance as a guiding precept, have had to deal with reactivity as much as their experimental colleagues in other social sciences.

The framework developed here thus seems to provide a promising route to finding out what makes results like those of the DG and similar games, particulary debatable. We restrict our analysis to the conceptual and theoretical clarification of the phenomenon through an interventionist framework, rather than devote this piece to the particular methodological analysis of a given design. We contend, however, that our conceptual contribution can be valuable to future applied research.

6 Conclusions

In social scientific experiments, the putative causes tested by the interventions come embedded in experimental scripts, rather than in pills, and thus operate through mental causation and social meaning, where this meaning is interpreted holistically. Experimental scripts embodying the treatment often give rise to some type of reactivity, whereby subjects modify their behavior as a result of some characteristics of the intervention, other than those related to the variable of interest. Whether this reactivity is suspected to be benign (if it does not have an effect on the relevant dependent variable) or malignant (if it does) will depend on the way those particular experimental scripts are processed and conceived by subjects. Moreover, this reactivity can, sometimes, be unique to each intervention, and thus, inseparable from each experimental script when the difference in outcomes between the control and the treatment group cannot guarantee that results are net of reactivity related input.

When we contrast the output of the treatment intervention with the control intervention (as in a modified DG versus a standard DG) in order to draw causal conclusions, we are implicitly assuming that the reactivity generated by each experimental script is benign or that, if it is malignant, it is equivalent across treatments (i.e., not idiosyncratic). While any of these assumptions may be true for any given intervention, they may not always hold in all cases. By stressing the need to specify the conditions under which these assumptions can hold, our analysis aims to contribute

to the debate over the limits of social scientific experimentation and specifically, about the validity of causal inferences generated by social experiments like the DG.

Ultimately, our intuitions about reactivity hinge upon, but also affect one's methodological position in the debate regarding the powers and the limits of social scientific experimentation. Indeed, while reactivity is traditionally considered by some a problem that can be either prevented by the use of control groups, or accounted for in the interpretation of results, it has represented for others a definitive obstacle to the mere possibility of investigating the social world experimentally (Harré and Secord 1972). A tension has traditionally existed between two seemingly irreconcilable views on the relationship between experimentation and the issue of reactivity: the experiment seen as the best environment to create the type of control that is needed to separate behavior into some of its relevant causal components, and. the view that experimentation is severely hindered by the fact that all social reality, including the experimental site is a thick, layered environment charged with social meaning, where that social meaning can only be interpreted holistically. Here we try to show that although reactivity is very likely a constitutive part of social experimentation, it is often benign. When it is not, it is often solvable through the standard practice of including a control group. Yet, we have also shown that when reactivity is not benign and is idiosyncratic, then it does pose problems to the inferential import of experiments. We have offered a conceptual framework to understand reactivity and argue that elucidating this concept provides a useful groundwork upon which we can build more nuanced, methodologically driven, case by case analyses.

Acknowledgements This paper was first presented at the Departamental Seminar in Uned, May 2016, the PSA 2016 Inem session, in Atlanta, and several other venues thereafter (including a Workshop on Reactivity organized at Bergen University in March 2020). I would like to thank the participants of these sessions for input and encouragement, as well as the reviewers and editors of EJPS.

Funding MECABIOSOC. FFI2017-89639-P. Ministerio de Ciencia, Innovación y Universidades.

Declarations

Ethical approval non applicable.

Informed consent non applicable.

Conflict of interest none.

References

Adair, J. G. (1984). The Hawthorne effect: a reconsideration of the methodological artifact. *Journal of Applied Psychology, 69,* 334–345.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics, 11*(2), 122–133.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton: Princeton University Press.

Campbell, J. (2007). An interventionist approach to causation in psychology. *Causal learning: Psychology, philosophy, and computation*, 58-66.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67–80.

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science, 74*(5), 981–995.

Guala, F., & Mittone, L. (2010). Paradigmatic experiments: the dictator game. *The Journal of Socio-Economics, 39*(5), 578–584.

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human behavior, 26*(3), 245–256.

Harré, R., & Secord, P. F. (1972). The explanation of social behaviour.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383–403.

Jimenez-Buedo, M. (2015). The last dictator game? Dominance, reactivity, and the methodological artefact in experimental economics. *International Studies in the Philosophy of Science, 29*(3), 295–310.

Jimenez-Buedo, M., & Guala, F. (2016). Artificiality, reactivity, and demand effects in experimental economics. *Philosophy of the Social Sciences, 46*(1), 3–23.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153–174.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*, 482–492.

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their Implications. *American Psychologist, 17,* 776–783.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in Behavioral Research* (pp. 143–179). New York: Academic Press.

Rosenthal, R. (1964). Experimenter outcome-orientation and the results of the psychological experiment. *Psychological Bulletin, 61,* 405.

Rosenthal, R. (1968). On the social psychology of the psychological experiment: 1, 2 the experimenter's hypothesis as unintended determinant of experimental results. *American Scientist, 51*(2), 268–283.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). Causation, prediction, and search. MIT press.

Teira, D., & Reiss, J. (2013). Blinding and the non-interference assumption in field experiments. *Philosophy of the Social Sciences, 43*(3), 358–372.

Teira, D. (2019). Placebo trials without mechanisms: How far can we go? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 77,* 101177.

Woodward, J. (2003). *Making things happen: A causal theory of explanation*. Oxford: Oxford University Press.

Woodward, J. (2007). Causation with a human Face. In Price and Corry (Eds.), *Causation and the constitution of reality*. Oxford University Press: 66–105

Woodward, J. (2008). Invariance, modularity, and all that. In S. Hartman, C. Hoefer, & L. Bovens (Eds.), *Nancy cartwright's philosophy of science* (pp. 198–237). Taylor & Francis.

Woodward, J. (2015). Methodology, ontology, and interventionism. *Synthese, 192*(11), 3577–3599.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*(1), 75–98.