Artificiality, Reactivity, and Demand Effects in Experimental Economics[*]

Maria Jimenez-Buedo and Francesco Guala

Date received: July 17th 2015

Contact address: mjbuedo@fsof.uned.es

Abstract:

A series of recent debates in experimental economics have associated demand effects with the artificiality of the experimental setting and have linked it to the problem of external validity. In this paper we argue that these associations can be misleading, partly because of the ambiguity with which "artificiality" has been defined, but also because demand effects and external validity are related in complex ways. We argue that artificiality (understood as unfamiliarity of the experimental environment) may be directly as well as inversely correlated with demand effects. We also distinguish between the demand effects of experimentation and the reactions that they may trigger and that might endanger experimental validity. We conclude that economists should pay more attention to the way in which subjects construe the experimental task, and learn to exploit subjects' reactivity to expectations in their experiments.

1 The problem of artificiality

The argument that the artificiality of the laboratory setting constitutes a serious methodological problem has a distinguished history in experimental economics. A classic statement can be found in a commentary written by Allen Wallis and Milton Friedman on one of the earliest laboratory experiments on demand theory.[1] Wallis and Friedman argued that

> It is questionable whether a subject in so artificial an experimental situation could know what choices he would make in an economic situation; not knowing it is almost inevitable that he

[*] The research for this paper was supported by grants FFI2011-28835 and FFI2014-57258-P of the Spanish Ministry of Science. We are grateful to two anonymous referees for their comments on a previous draft.

[1] The experiments were reported in Thurstone (1931). On early experiments in demand theory, see Moscati (2007).

> would, in entire good faith, systematize his answers in such a way as to produce plausible but spurious results (Wallis and Friedman 1942: 179).

The Wallis-Friedman argument struck a chord, fuelling economists' scepticism about laboratory experiments for decades. In the 1970s and 1980s, pioneers like Vernon Smith (1982) and Charles Plott (1991) responded to this scepticism arguing that experiments are mainly valuable as tests of economic theories, and any theory that has been refuted in the laboratory is unsatisfactory because it fails to be universally valid.[2] The appeal of this argument, however, is limited in a discipline that, like economics, does not aim at capturing universal laws of nature.[3] Thus unsurprisingly, in spite of the success of experimental economics and its integration within the discipline, the artificiality worry has not disappeared. According to Arthur Schram, for example,

> A major obstacle to the external validity of an experiment is the artificiality of the setting. If the laboratory institutions and incentives do not sufficiently mirror those of the outside-the-laboratory situation they intend to study, the loss of external validity may be significant (Schram 2005: 226).

"Artificiality" is a vague notion. Although sometimes it is used to refer to the abstract nature of the experimental task, it is often used to refer to anything that is different in the experiment and its target in the "real world" (see Schram as an example), and is seen as putting validity at risk.[4] Validity, strictly speaking, is the property of an inference from experimental evidence to a claim or hypothesis. According to a standard definition, an inference is *internally valid* when it is correct and the hypothesis concerns mechanisms and causes that are at work within the laboratory setting; it is *externally valid* when it is correct and the hypothesis refers to causes and mechanisms that are operative in some non-laboratory situation of interest.[5] The artificiality worry thus concerns a

---

[2] See also Wilde (1981: 143), Loomes (1989: 173), Hey (1991: 10).

[3] See e.g. Guala (2005: 147-160), Levitt and List (2007: 153-4).

[4] See also Berkowitz and Donnerstein (1982), Mook (1983), Starmer (1999), Lucas (2003), Bardsley (2005), Bardsley et al (2010).

[5] Cf. Guala (2005), Bardsley (2005), Jimenez-Buedo and Miller (2010), Jones (2011).

specific way in which experimental data may lead us astray, or in which we could make an invalid inference from data to non-experimental causes. The worry is that experimenters may be studying tasks that are *abstract and unfamiliar to participants*. The behaviour of participants thus may be influenced by causal factors that are quite different from those that would be operative in other (more familiar) circumstances. An inference from data to non-laboratory causes could be seriously misguided, leading to identify decision processes that are of little relevance in most non-laboratory situations.

Artificiality lately has also been associated with so-called "experimental demand effects", or the behavioural changes that may be prompted by subjects' awareness of being under study. Steve Levitt and John List for instance point out that

> humans, unlike Galileo's rolling balls or Uranium239, know that they are participating in experiments. Making decisions in an artificial environment and general awareness of the fact that their actions are being observed and recorded might influence how people behave (2005: 5).[6]

Again, artificiality and subjects' awareness are considered a threat to the validity of experimental results. The worry is that the behaviour observed in some laboratory experiments might tell us more about subjects' reactions to experimenters' demands, than about their motives when they face similar tasks in the "real world". As an antidote, Levitt and List claim, economists should design field experiments in which subjects face familiar tasks and are unaware of being studied.

In this paper we will examine critically the concept of artificiality and the way it has been used in the experimental literature. For ease of exposition, in the next section we introduce the Dictator

[6] The paragraph that we are quoting has disappeared during the transition from working paper (Levitt and List 2005) to published article (Levit and List 2007). We suspect that the cut was made for reasons of space, since the spirit of the published text is very much the same, and there is no reason to believe that the authors changed their mind on this matter.

Game, a paradigmatic experimental design which has been widely criticized for its artificiality but which nevertheless has enjoyed huge success over the last two decades. The Dictator Game will be used as an example and will help illustrate concretely some of the general methodological issues discussed in this paper. Section 3 addresses the relation between artificiality and demand effects. Our thesis is that, under a common interpretation of the term, the artificiality of the experimental setting and the effects that the awareness of participating in an experiment can have on the behaviour of subjects are related in complex ways, and that there is no reason to believe that – in general – artificiality or awareness affect the validity of experiments univocally. The subsequent section (section 4) offers an analysis of the broader phenomenon of reactivity and the conditions under which it can lead to experimental artefacts. Finally (sections 5 and 6), we will argue that the key to make valid inferences from experimental data lies not so much in eliminating demand effects as in retaining control over them, and understanding subjects' reactions to expectations is crucial to achieve this goal.

2 An example: the Dictator Game

Imagine you are participating in an experiment: you are sitting in front of a computer terminal, surrounded by partitions that prevent you from seeing the other participants. A set of instructions describe the task that you will perform in the next few minutes. First, you will be matched randomly with another subject. Her identity will remain unknown to you, and your identity will be unknown to her. Then, you will be assigned different roles, labelled A and B. All B-participants will wait for A-participants to make a decision. A-participants will see a small box appear in the middle of their computer screens, preceded by the following message:

> You have 20 euro in your account. You now have the opportunity to transfer some of this money to the account of a B-participant who has been matched randomly with you. How much

money would you like to transfer to the other participant? Write in the box a number from 0 to 20 and press Continue.

When A has made her decision, the experiment will end. At that point you will be given a sealed envelope with a participation fee of five euro plus the sum of money that is in your account. The instructions emphasize that your identity and the identity of the other participants will remain secret during and after the experiment, and that no one else will know what you have decided to do or how much money you have earned in the experiment.

The setting we have just described is known as the Dictator Game, and is one of the best known designs in experimental social science.[7] It has been replicated hundreds of times by economists, and over the last ten years it has also become increasingly popular in psychology, biology, and anthropology. In spite of its success, however, the Dictator Game is a controversial design. The results consistently show that roughly half of the dictators[8] depart from the profit maximizing strategy and choose to give some money to the recipients, the mean allocation being 20% of the initial endowment. Moreover, a consistent minority of dictators choose to split the sum in two equal parts (cf. Camerer 2003).

It is not clear how these results should be explained. According to some social scientists, behaviour in the Dictator Game shows that people are not the selfish income-maximizers postulated in many economic models: they are willing to benefit other individuals even if it is costly, and their preferences have a "social" or "other-regarding" element.[9] Others disagree: they claim that the behaviour observed in Dictator Games cannot be used to draw inferences about people's preferences, because it is an "experimental artefact". The choices made by dictators are caused by

[7] Kahneman et al (1986) and Forsythe et al (1994) are generally considered the seminal papers on the Dictator Game. For a meta-analysis see Engel (2011), and for a methodological overview see Guala and Mittone (2010).

[8] The dictators are "A-subjects", in the terminology above. Experimental instructions avoid loaded terms such as "dictator" and try to use a terminology that is as neutral as possible. B-subjects are known as "recipients" in the scientific literature.

[9] See e.g. Andreoni and Miller (2002), Fehr and Fischbacher (2002).

the peculiar circumstances in which the subjects are artificially placed (e.g. Levitt and List 2005, 2007; List 2007, Bardsley 2008).

The sceptics seem to have a point: in spite of its simplicity, the Dictator Game is a really odd situation. Although we always have the opportunity to give money to strangers, we rarely receive a sum especially for this purpose (from another stranger!); and we are rarely told that there is an anonymous individual with whom we could share some of this windfall money. The hypothesis that subjects find the situation perplexing therefore is not far-fetched. The "oddness" of the situation is due to two separate causes: on the one hand, the unfamiliarity of the environment; on the other, the fact that the subjects are under experimental scrutiny. Both features, moreover, are potentially problematic because human subjects are aware of the situation they are in: awareness may induce subjects to comply with the "demands" of the experimenter, invalidating the inferences that the latter will make.

In the course of the paper we will refer frequently to Dictator Game experiments to flesh out some interesting aspect of the artificiality problem. As we will do this, it is important to keep in mind that the Dictator Game is a special case for the reasons mentioned above. Nevertheless, we think that it is a useful case partly in virtue of its oddity, which explains why it has been extensively chosen as a case study to illustrate the problems of artificiality and demand effects.

3 Task construal and demand effects of experimentation

We owe the notion of demand effects of experimentation to the classic work of Martin Orne, the first psychologist who tried to study systematically the experimental situation as a social phenomenon. Orne investigated how participants, in their attempt to interpret the experimental context, try to guess the experiment's purpose and (often unconsciously) direct their behavior to fit

that interpretation. As part of his theoretical contribution, he coined the notion of "demand characteristics of experimentation" (Orne, 1962, 1969) to refer to the set of cues, instructions, and interactions that define any experiment and that subjects use to direct their behaviour. The crucial idea behind Orne's concept is that a number of cues and instructions are not explicitly designed but instead emerge spontaneously from the interaction between subject and experimenter. These cues, therefore, may not be under direct experimental control. Their behavioral consequences are the *demand effects of experimentation*.

A crucial source of demand effects is *task construal*: in experiments that provide unfamiliar scenarios subjects look for cues that facilitate the interpretation of the task. According to a radical interpretation, all experimental instructions are necessarily incomplete; subjects must always evaluate the circumstances and fill the gaps using whatever environmental cues and previous experiences that share some elements with the current situation. They, in a sense, construe "their own" experimental task.

Although experimental cues may be unintended, their behavioural effects can be studied empirically, at least in principle. This idea has a long standing tradition in social psychology, and recently has received attention in experimental economics as well. In the case of the Dictator Game, for example, it is possible to manipulate the basic design in such a way as to send different cues about the purpose of the task, while maintaining the same structure of economic incentives. Bardsley (2008) for instance has designed an experiment in which dictators, in addition to giving, can choose to take money from recipients. In all of his treatments a significant number of dictators choose to take from, rather than give money to, their experimental counterparts. Bardsley interprets this as an artefact of experimentation: both in his game and in the standard Dictator Game subjects are reacting to the cues that the protocol supplies about what constitutes "appropriate" behaviour. When the experiment (the standard Dictator Game) seems to be about giving, subjects give. When the experiment (Bardsley's Dictator Game variant) seems to be about taking, subjects feel free to take

from recipients.[10]

The notions of demand effect and task construal are useful tools to analyse experiments like the Dictator Game. Both have the advantage of emphasising the active role played by subjects in the interpretation of experiments. To the extent that "artificiality" can create methodological difficulties, these are not based on the fact that the task is abstract or unfamiliar, but on the rather general problem that, as any other element in the setting, it can prompt reactions that depend on subjects' uncontrolled interpretations of the task. The problem is, though, as we shall see, that familiar tasks can also give rise to uncontrolled interpretations on the part of subjects. Since the subjects engage with tacit and explicit instructions provided by the experimenters, the latter obviously play some role in inducing the behaviour that is observed in the laboratory. But it would be fallacious to consider the behaviour as determined or "implied" by the experimental task only.

Daniel Zizzo (2010) has tried to construct a comprehensive framework for the analysis of the interaction between the task and its interpretation using the concept of "experimenter's demand effect". Zizzo defines these effects as

> changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior (behavior 'demanded' from them). (2010: 75)

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.

[10] For similar effects, see also Dana et al (2007), List (2007), Zizzo and Fleming (2014).

(3) Positively correlated expected and true objectives.

Zizzo argues that in the first case, in which subjects' expectations regarding the goals of the experiment are orthogonal to the real objectives, demand effects do not pose a threat to the validity of experimental inference. The second case, in which subjects' guesses are opposite to the experimental objectives, does not constitute a problem for validity either: whatever effect we observe we can attribute to the treatment, minus the effect of the experimenter's demand.[11] Only the third case is truly problematic according to Zizzo: 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 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 framework emphasises the question of whether the experiment provides cues to subjects about its purpose. The formulation in terms of correlation between true and expected objectives, however, is not entirely satisfactory.

Assuming for the sake of the argument that the expectations of the experimenter can be identified correctly,[12] it is not clear what behavioural implications such a correlation would have. Zizzo says that "correlation" means "correlation between *actions implied* by the expected objectives and *actions implied* by the true experiment objectives" (2011: 86, n. 16). For example, an experiment where subjects have the opportunity to give money – as in the standard dictator's game – would

[11] It should be noticed, however, that there may be "false negatives" – the treatment may seem to have no effect, where in fact it has been counterbalanced by the experimenter's demand. And even in case of a positive effect, its size could be systematically underestimated.

[12] One problem is that the relationship between the topic that the experimenter is studying and the intentions of the experimenter may be rather loose. For example, in some between-subjects experiment it can be virtually impossible to guess what the experimental manipulation is, and hence to infer what the "appropriate" or "expected" behaviour might be. Experiments are run "between subjects" when different treatments are administered to different groups, and "within subjects" when the same sample of subjects is administered all the treatments. In the first case, each subject has the opportunity to observe only part of the experiment and therefore cannot reliably infer what the hypothesis under test might be.

naturally imply the action "give". But notice that the goals of the experimenter are only one side of the coin, and by themselves do not have any specific behavioural consequences: the motives of the subjects play a crucial role as well. In order to figure out whether the behaviour of the subjects is likely to change in response to the experimental stimulus, and how, we must pay attention to the way in which subjects *react* to the perceived experimental goals.

4 Reactivity and control

Following a terminology that is common in social psychology, we shall call *reactivity* the phenomenon that occurs when individuals alter their behaviour because of the awareness of being studied.[13] Reactivity is intimately related to the notion of demand effect of experimentation, but we argue that it has some conceptual advantages that make it more suitable to methodological theorizing. One important advantage is its neutrality regarding the behavioural outcome of the task. While the term "demand effect" is often used to refer to the phenomenon of active task construal, it is also (more often than not) used to refer to the biases that task construal may introduce in the experimenters' interpretation of results. Reactivity has less of a negative connotation, and allows us to identify the phenomenon independently of whether it creates methodological problems or not.

Reactivity can take many forms, which may have different implications for the validity of experimental inferences. Here we focus on three reactive mechanisms that may in principle operate alternatively, or in conjunction:

(i) *Puzzle solving*: when confronted with the experimental situation, subjects tend to conceive of their role as one in which a puzzle must be solved. The puzzle may consist in finding out what one is supposed to do as subject, or what the experiment really is about. In addition, especially in the case

[13] See e.g. Adair (1984).

of experimental economics, some experiments look like abstract mathematical puzzles. One way to comply with the experimenter's demand, then, is to look for the optimal solution, where the optimum is defined by some logical or quasi-logical criterion. Clearly the search for an optimal solution requires that subjects form an opinion on the goal of the experiment (task construal) and thus, if the perceived objective coincides with the real objective, we may have an example of Zizzo's third case.

(ii) *Cooperative attitude*: subjects may be motivated to solve a puzzle intrinsically (they like to solve puzzles) or by means of incentives (monetary, or otherwise): in the latter case they are motivated by a prize. But a further plausible motivation may be their desire to act according to what they think they are expected to do.[14] To satisfy this motivation requires that subjects identify the goal of the experiment, and correctly infer what the experimenter would like to observe. Again, we may have an example of Zizzo's third case.

(iii) *Evaluation apprehension*: why would subjects want to give the experimenter the data that he is looking for? The simplest answer is that they want to make him happy. But another powerful motivation may be a desire to *look good* in the eyes of the experimenter. The demand effect in this case would not be triggered by subjects' wish to attain what they think are the experimenter's goals, but by their desire to convey a given image of themselves to the experimenter. In this more complex or strategic interaction the subject tries to guess the experiment's goals and chooses a behavioural pattern that he or she thinks will project a positive image.

Notice that subjects' motivations become increasingly complex or layered as we move from (i) to (iii). The first mechanism only requires that subjects are motivated to find the "right" solution to the task; the second requires that they do it to please the experimenter; the third requires that they be

[14] An opposite motivation may be the desire to disrupt these expectations, although psychologists tend to assume that this happens only rarely.

concerned with what the experimenter may infer about their skills or moral character, and that they act strategically in order to control that inference.

Notice also that reactivity does not need to constitute, in itself, a source of experimental bias. A research artefact is normally defined as a systematic bias, *uncontrolled and unintentional*, that can threaten the validity (internal or external) of one's conclusions (Strohmetz and Rosnow 2004). Thus, the reactivity of experimental subjects must be unregistered by the experimenter in order to constitute a research artefact. The relevant literature often displays a fair degree of confusion in this respect: all too often "artefacts" are used indistinctly (a) to refer to the phenomenon whereby subjects alter their behaviour due to the awareness of being studied, and (b) to refer to the inferential mistakes that this may cause.

Another point worth stressing is that there is little reason to think that artificial designs are *more* likely to cause artefacts associated to reactivity or to demand characteristics: for example, experimenters are not in principle more likely to send uncontrolled cues that hint at the experimental purposes, just because the experimental situation is abstract or unfamiliar to subjects. On the contrary, an unfamiliar setting could be exploited to send more controlled cues to participants about the purposes of the experiment or the kind of behaviour that is expected. In a rich environment, subjects may react to a variety of stimuli, and the way in which they construe the task may become very difficult to predict (which is, incidentally, one of the reasons why experimental economists have traditionally privileged abstract designs). Therefore, an artificial setting does not seem to be, prima facie, more problematic in terms of the reactions it may trigger and their correct interpretations by researchers. It is perhaps important to stress once more that whatever cues the setting provides, they are to be considered problematic only in so far as they are sent *inadvertently* by the experimenter, i.e., only if the experimenter does not take them into account when he or she is interpreting subjects' behaviour as a response to the treatment.

In sum, an artificial experimental environment per se does not need to make it more difficult for the experimenters to make valid inferences about the motives of subjects. The elements that can bias experimental results or cause researchers to make invalid inferences pertain to each individual design (and its relation to the background knowledge that motivates this design), but this seems to be orthogonal to the question of whether experimental designs involve familiar or unfamiliar tasks.

Similarly, it is not clear how the artificiality or unfamiliarity of a given experimental setting may enhance the zealousness with which subjects obey instructions, the apprehension with which they face evaluation, or their reactions to the authority of experimenters. One could say, by the same token, that the more natural or recognizable the experimental task or situation, the more subjects may wonder about the researchers' hidden motives, and the more zealousness, apprehension, or obedience/rebelliousness they will display. Human subjects may change their behaviour because they are under observation or scrutiny, regardless of whether they are engaged in a familiar or in an artificial experimental situation (think for example of the difficulty for people to display spontaneous patterns of behaviour when natural, physiological acts are being studied).

The last point we would like to highlight is that whether or not reactive behaviour has the potential to bring about experimental bias seems to be unrelated, in principle, to whether subjects correctly guess the nature of the experimenter's goals. Take mechanism (i) above, whereby subjects tend to conceive of the experimental context as containing a puzzle that they have to decipher. While Zizzo's account suggests that a correlation between the expected and true objectives of an experiment is problematic, a perfect coincidence or matching between them can also be an antidote against bias: think of those experiments in which subjects are explicitly required and properly incentivised to perform a given task at their best level of their capacities. In these cases, perceived and real objectives of the experiments coincide, but there is no room for biasing effects, because subjects pursue the goals that the experimenter wants them to pursue (their best performance).

5 Expectations and norms

We have argued that reactivity provides a powerful framework to analyse the problems that may be posed by "artificial" experiments. First, it allows us to separate the different mechanisms (both cognitive and motivational) of "demand effects", and second, it provides a neutral terminology that does not ipso facto link subjects' reactions to invalid inferences drawn from the experiment. Finally, we have pointed out that reactivity is a problem only when it goes undetected. In the rest of the paper we will argue that in order to avoid validity problems it is necessary to understand the mechanics of reactivity. What do subjects react to, exactly?

One possible explanation of the behaviour observed in the standard Dictator Game and similar experiments, which has been extensively discussed in the literature, is that people care about the earnings of others. They may care about it directly – if they simply want to increase other people's welfare – or indirectly – if they implement abstract principles of fairness concerning payoff distribution.[15] If this is the case then the same results should be observed whenever the same distributional choices are being made, irrespective of how subjects construe the experimental task.

There is extensive evidence however that subjects behave differently depending on whether their choices are known to others. In games where the payoffs are expressed in "chips" with different monetary values to each player, for example, many subjects are happy to simply pretend to be fair (by sharing the chips, but not the money, equally) if the other players do not know how much the chips are worth (Kagel et al. 1996). Other kinds of uncertainty – regarding, for example, the dictator's responsibility for the outcome – also create a "moral wiggle room" that is exploited by experimental subjects for selfish purposes (Dana et al. 2007).

[15] For a survey of the (large) literature on so-called "social preferences", see for example Fehr and Fischbacher (2002) and Fehr and Schmidt (2006).

So it seems that people often care about others' opinions, more than (or in addition to) others' welfare. Subjects try to anticipate the expectations of a relevant audience, and adjust their behaviour in such a way as to not disappoint the audience. The audience in experiments like the Dictator Game is partly constituted by other subjects, if the latter know the distribution of payoffs. But it is also constituted by the experimenter, who is almost invariably informed (or has the means to find out) about subjects' behaviour. In the latter case – and if subjects care about the expectations of the experimenter – there is clearly a potential for reactivity playing a role in subjects' behavior.

Note however that, despite what is often assumed in methodological discussions, the presence of an audience in itself is far from unusual or "artificial". In many real-world situations our actions do take place in front of friends, colleagues, or just occasional bystanders, and accordingly our actions are influenced by what we think are their expectations about our behaviour. So the fact that subjects react to the expectations of a given audience should not necessarily be conceived as a threat to external validity. Notice also that artificiality and expectations are related in complex ways. There is a sense in which the artificiality of an experimental setting may actually *hamper* subjects' reaction to other people's expectations, because an unfamiliar situation can make it *harder* for the subjects to figure out what kind of expectations the audience might have. Those experimenters who have bothered asking their subjects what behaviour they think is "appropriate" in the Dictator Game, for example, have found that the answers vary enormously and that a large proportion of subjects simply admit that they do not know (Bicchieri 2006: 126).

These data suggest that what is potentially problematic about the interpretation of the standard Dictator Game is not so much "artificiality" or subjects' reaction to an experimental "demand", but that we do not fully understand how the task is construed by the experimental subjects. This in turn suggests that in behavioural experiments like the Dictator Game one can learn a lot by manipulating the cues so as to change the construal, and hence the perceived expectations. Such manipulations may in principle make the experimental environment either more or less artificial (in the sense of

more or less similar to a familiar situation), but the "artificiality" need not have a direct impact on the degree of control that experimenters have over subjects' expectations. Elements that can be manipulated in the Dictator Game include, for example, the source of subjects' endowments ("windfall" money) and the identity of the recipient (a randomly selected experimental subject). Such manipulation can have a tremendous impact on subjects' behaviour (regardless of whether they make the experiment more or less familiar): Cherry et al. (2002) for example have observed that 95% of subjects donate nothing in a Dictator Game when the dictators earn the money by answering the questions of a GMAT quiz correctly. In another experiment, over 73% of the dictators gives money when the recipient is identified with a "reputable charity" like the Red Cross, and the average level of donations is tripled compared to an anonymous recipient condition (Eckel and Grossman, 1996).[16]

How can we explain this evidence? Bicchieri (2006), Levitt and List (2007) and Smith (2008) have argued that the key is to understand the complex nature of norm-driven behaviour. Subjects' behaviour (and a fortiori, experimenter's demand effects) may be guided by perceived normative expectations. But far from being an anomaly of "artificial" experiments, people's sensitivity to norms is an extremely common and important behavioural phenomenon. The subject who does not share the money earned in a GMAT task is reacting to a familiar norm that assigns property rights over the resources that she has produced with her own labour. Similarly, the subject who gives money to the Red Cross reacts to a norm that prescribes to help people in need. And in both cases, the norm is likely to be shared with the experimenter and with the other subjects. An inference from lab to world would be externally valid, in these circumstances, *because* the subjects are reacting to the expectations of an audience (not in spite of it), given that the experimental audience is representative of other audiences found in real-life situations.

[16] For other similar examples see e.g. Branas-Garza (2007) on the effect of making the passive role of the recipient salient; Krupka and Weber's (2013) study of "giving" and "taking" frames; and Jakiela's (2015) finding that subjects in rural Kenia react to a "status" cue but not to an "effort" cue.

If the behaviour observed in experiments is explained by compliance with norms, then tasks like the Dictator Game may be perfectly appropriate tools for the investigation of social behaviour. In order for this to be the case, however, the experimenter must achieve control on the main determinants of behaviour: she must understand what sort of norms may be triggered by what sort of cues, and what kind of audience is important for the subjects. When Cherry and his colleagues introduced the GMAT test, they were aware that they were manipulating a norm of asset legitimacy or private property. Similarly, Eckel and Grossman knew what they were doing when they indicated the Red Cross as the recipient of their Dictator Game. But experimental economists in general are not always aware of the methodological implications of experimenting with social norms. Such methodological awareness can be attained only if we shift the focus of attention from the control of individual preferences to the control of social norms.

6 Controlling norms

What is a norm, then, and how can it be controlled in the laboratory? When we say that "you ought to do" something, we usually intend that we expect you to do it even if you may have some reason to do otherwise. For example, if we say that you ought to be here at noon, we expect you to do it even if you have some reason to delay. If you tell us that you are going to be late because you want to take a nap in the park, we will probably not take it as a legitimate justification. We will be annoyed, and we will form a bad opinion of you.[17]

A remarkable feature of norm-driven behaviour is this: the fact that you are expected to do something may also be a reason for you to do it. Cristina Bicchieri has proposed a definition of

[17] This obviously holds for some reasons only: if you are late because you have been hit by a truck and required some medication, we will consider it a good justification. Which deviations from a norm are acceptable depends on the relative cost of compliance and the strength of the norm.

social norm that is able to capture this conditionality on expectations, and that we will use in this paper. A rule R, according to Bicchieri (2006), is a *social norm* in a population P if

> (a) the members of the group believe that a sufficiently large subset of P conforms to R in situations of type S, and either
> (b) they believe that a sufficiently large subset of P expects them to conform to R in situations of type S; or
> (b′) they believe that a sufficiently large subset of P expects them to conform to R in situations of type S, prefers them to conform, and may punish deviations from R.

Notice that external reasons to conform (i.e. sanctions) are mentioned explicitly only in condition b′. When there is no expectation of punishment, therefore, non-conformity with expectations generates motivations that are *internal* to the decision-maker, for example in the form of a desire to please the audience.

Bicchieri argues explicitly that norm-compliance is conditional on expectations, and that expectations are relative both to a population P and to a situation S. The so-called problem of "artificiality" then is often the problem of inferring R from S, a particularly tricky task when the subjects are unable to associate the experimental task with a familiar real-world task. But it is important to realize that "artificiality" and expectations may interact in various ways: on the one hand, because an unfamiliar design often sends fewer cues to the experimental subjects, it is arguably less likely to send cues *inadvertently* – that is, to create expectations that are uncontrolled by the experimenter. On the other hand, in the absence of a familiar cue the subjects may engage in wild speculations about the goal of the experiment or the expectations of the audience. A small detail of the design may be used by subjects to construe an idiosyncratic interpretation of the task, and in such cases there is a risk that the experimenter may lose control of the normative elements of the situation.

The behaviour triggered by experimental cues is in a very obvious and straightforward sense an experimenter's demand effect in Zizzo's sense. But the fact that norm-compliance has been induced by the experimenter ("artificially", inevitably) does not necessarily invalidate the inferences that we make from the data. Quite the contrary: just as it makes good sense to induce preferences in some experimental circumstances, so it makes good sense to induce norm-driven behaviour in other experimental contexts. The important point is that the experimenter must retain full control over the experimental procedures – or, in other words, that the inferences drawn from the experimental data must not be confounded by any unintended effect of the experimental design. To retain experimental control the experimenter must be aware of the potential effects of any cues that may be implicit or explicit in the experimental design. She must be aware of the way in which a certain task may be classified, and of the possible association between the classification and any behavioural rule R. This in turn is possible only if the experimenter is acquainted with the cultural beliefs and expectations of the population from which the experimental subjects and their audience have been drawn.

Some fascinating examples come from the experiments performed by anthropologists with populations that share different cultural norms from those that are prevalent in Western societies. Lesorogol (2007) for example reports that the Samburu, a group of nomadic pastoralists from Kenya, give the recipient roughly 40% of the endowment in a standard Dictator Game – an unusually generous behaviour compared to what is observed in Western countries. When the Dictator Game is framed as a "meat sharing" task, however, the modal offer decreases to 20%. During post-experimental interviews, the subjects explain that such a portion corresponds roughly to the size of a hind leg, which is considered the normatively appropriate donation when the carcass of a goat is shared among the Samburu.[18]

[18] See also Ensminger (2004), Cronk (2007), Wiessner (2009), Gerkey (2013), Barr et al. (2015), Jakiela (2015).

The methodology followed in this experiment is entirely appropriate, given the goals of the research. The meat sharing task is framed in such a way as to elicit normatively appropriate behaviour, and in this sense the results may be considered "demand effects". But the behaviour is not an "artefact" (quite the contrary) and the inferences are valid, because the experimenter retains control over the experimental manipulation. Validity depends on our capacity to make reliable inferences from the data, and the background knowledge of the experimenter (concerning P, S and R) is crucial in this respect.

7 Conclusions

The methodological literature in experimental economics tends to associate artificial designs with lack of external validity and demand effects. In this paper we have argued that these three concepts are entangled in a complex fashion. Part of the problem is that the term "artificiality", as it is used in the literature, refers to at least three different features of experiments, that ought to be kept distinct: (i) the fact that subjects are observed by an audience, (ii) the fact that they are placed in an unfamiliar situation and (iii) the fact that there is often only an approximate or analogical correspondence between the experimental setting and the target situation of interest. Once these three features are distinguished, we can see that there is no univocal relationship between artificiality, demand effects, and external validity; in some cases artificiality (understood as involving unfamiliar tasks or settings) can prevent biases associated to reactivity, since the key to retain experimental control is to understand the reaction of experimental subjects to the expectations of the other subjects and of the experimenter.

Emphasising the role of reactivity, expectations, and the audience, helps identify some misguided solutions to the problem of external validity. Levitt and List (2007) for example promote field experiments as an antidote to the artificiality of designs such as the Dictator Game. But clearly we do not make expectations and audiences disappear just by running a field experiment. If there are

norms (and there often are) we must be aware of how the experimental manipulation may trigger cues and hence norm-driven behaviour. If we lose control over this aspect, we are going to derive invalid inferences regardless of whether the experiment is run in the laboratory or in the field. (In the field the consequences may actually be even worse, because we might feel justified in drawing policy conclusions that are in fact unwarranted.)

As an example of valid inferences drawn from experimental data, we have mentioned the studies of social norms done by economists and anthropologists using "artificial" designs like the Dictator Game. Such studies do produce valid results when the experimenters retain control over the reactions of their subjects to the experimental cues. One important caveat is that the results are typically limited to a particular culture (meat sharing norms are not common among Western students, for example). This is the sort of validity that can be attained; but surely a valid local inference is more valuable than a mistaken universal inference, or than no inference at all.

REFERENCES

Adair, J. G. (1984) "The Hawthorne Effect: A Reconsideration of the Methodological Artifact", *Journal of Applied Psychology*: 69: 334-345.

Andreoni, J. and Miller, J. (2002) "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism", *Econometrica* 70: 737-753.

Bardsley, N. (2005) "Experimental Economics and the Artificiality of Alteration", *Journal of Economic Methodology* 12: 239-251.

Bardsley, N. (2008) "Dictator Game Giving: Altruism or Artifact?" *Experimental Economics* 11: 122-133.

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., and Sugden, R. (2009) *Experimental Economics: Rethinking the Rules*. Princeton University Press.

Barr, A., J. Burns, L. Miller, and I. Shaw (2015) "Economic Status and Acknowledgement of Earned Entitlement", *Journal of Economic Behavior and Organization* 118: 55-68.

Berkowitz, L. and Donnerstein, E. (1982) "External Validity Is More than Skin Deep: Some Answers to Criticisms of Laboratory Experiments", *American Psychologist* 37: 245-257.

Bicchieri, C. (2006) *The Grammar of Society*. Cambridge University Press.

Branas Garza, P. (2006) "Poverty in Dictator Games: Awakening Solidarity", *Journal of Economic Behavior and Organization* 60: 306–320.

Branas Garza, P. (2007) "Promoting Helping Behavior with Framing in Dictator Games", *Journal of Economic Psychology* 28: 477-486.

Camerer, C. (2003). *Behavioral Game Theory*. Princeton University Press.

Cherry, T.L., Frykblom, P., Shogren, J.F. (2002) "Hardnose the Dictator", *American Economic Review* 92: 1218–1221.

Chiesa, M. and S. Hobbs (2008) "Making Sense of Social Research: How useful is the Hawthorne effect?" *European Journal of Social Psychology* 38: 67-74.

Cronk, L. (2007) "The Influence of Cultural Framing on Play in the Trust Game: a Maasai Example", *Evolution and Human Behavior* 28: 352-358.

Dana, J., R. Weber, and J. K. Kuang (2007) "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness". *Economic Theory* 33: 67-80.

Eckel, C., Grossman, P.J. (1996) "Altruism in Anonymous Dictator Games", *Games and Economic Behavior* 16: 181–191.

Engel, C. (2011) "Dictator Games: A Meta Study", *Experimental Economics* 14: 583-610.

Ensminger, J. (2004) "Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa". In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr,

and H. Gintis (eds.) *Foundations of Human Sociality*. Oxford: Oxford University Press, pp. 356–381.

Fehr, E. and Fischbacher, U. (2002) "Why Social Preferences Matter – The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives", *Economic Journal* 112: C1–C33.

Fehr, E. e Schmidt, K. (2006) "The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories". In S. Kolm & J.M. Ythier (eds.) *Handbook of the Economics of Giving, Reciprocity and Altruism,* Amsterdam: Elsevier.

Fleming, P. and Zizzo, D. (2014) "A Simple Stress Test of Experimenter Demand Effects", *Theory and Decision,* online first, doi: 10.1007/s11238-014-9419-2.

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M. (1994) "Fairness in simple bargaining experiments". *Games and Economic Behavior* 6: 347–369.

Gerkey, D (2013) "Cooperation in Context: Public Goods Games and Post-Soviet Collectives in Kamchatka, Russia". *Current Anthropology* 54: 144-176.

Guala, F. (2005) *The Methodology of Experimental Economics*. Cambridge University Press.

Guala, F. and Mittone, L. (2010) "Paradigmatic experiments: The Dictator Game". *Journal of Socio-Economics* 39: 578-584.

Hey, J.D. (1991) *Experiments in Economics*. Oxford: Blackwell.

Jakiela, P. (2015) "How Fair Shares Compare: Experimental Evidence from Two Cultures", *Journal of Economic Behavior and Organization* 118: 40-54.

Jimenez-Buedo, M. and Miller, L.M. (2010) "Why a Trade-off? The Relationship between the External and Internal Validity of Experiments", *Theoria* 69: 301-321.

Jones, M.K. (2011) "External Validity and Libraries of Phenomena: A Critique of Francesco Guala's Methodology of Experimental Economics", *Economics and Philosophy* 27: 247-271.

Kahneman, D., Knetsch, J., Thaler, R. (1986) “Fairness and the Assumptions of Economics”. *Journal of Business* 59: S285–300.

Krupka, E.L. and Weber, R. (2013) “Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?”, *Journal of the European Economic Association* 11: 495-524.

Levitt, S. D. and List, J. A. (2005) “What Do Laboratory Experiments Tell Us About the Real World?”, NBER Working Paper.

Levitt, S. D. and List, J. A. (2007) “What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?” *Journal of Economic Perspectives* 21: 153-174.

Lesorogol, C. K. (2007) “Bringing Norms In: The Role of Context in Experimental Dictator Games”, *Current Anthropology* 48: 920-926.

List, J. A. (2007) “On the Interpretation of Giving in Dictator Games”. *Journal of Political Economy* 115: 482-493.

Loomes, G. (1989) “Experimental Economics,” in J. D. Hey (ed.) *Current Issues in Microeconomics*. New York: St. Martin’s Press, pp. 152-78.

Lucas, J.W. (2003) “Theory-Testing, Generalization, and the Problem of External Validity”, *Sociological Theory* 21: 236-253.

Mittone, L. and Ploner, M. (2012) “Asset Legitimacy and Perceived Equity in the Dictator Game”, *Journal of Behavioral Decision Making* 25: 135–142.

Mook, D.G. (1983) “In Defense of External Invalidity”, *American Psychologist* 38: 379-387.

Moscati, I. (2007) “Early Experiments in Consumer Demand Theory: 1930-1970”, *History of Political Economy* 39: 359-401.

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 and R. Rosnow (eds.) *Artifact in Behavioral Research*. Academic Press, 143-179.

Plott, C. R. (1991) “Will Economics Become an Experimental Science?,” *Southern Economic Journal* 57: 901–19.

Rosenthal, R. (1963). “On The Social Psychology Of The Psychological Experiment: 1, 2 The Experimenter’s Hypothesis As Unintended Determinant Of Experimental Results”. *American Scientist* 51: 268-283.

Ruffle, B. J. (1998) “More is Better, but Fair Is Fair: Tipping in Dictator and Ultimatum Games”, *Games and Economic Behavior* 23: 247–265.

Schram, A. (2005) “Artificiality: The Tension between Internal and External Validity in Economic Experiments”, *Journal of Economic Methodology* 12: 225-237.

Smith, V. L. (1982) “Microeconomic Systems as an Experimental Science,” *American Economic Review* 72: 923–55.

Smith, V.L. (2008) *Rationality in Economics: Contructive and Ecological Forms*. Cambridge University Press.

Starmer, C. (2005) “Normative Notions in Descriptive Dialogues”, *Journal of Economic Methodology*12: 277-89.

Strohmetz, D. and Rosnow, R. (2004) “Artifacts in Research Process”. In Lewis-Beck, M., Bryman, A. E. and Liao, T. F. (eds.) *The Sage Encyclopedia of Social Science Research Methods* (Vol. 1). Sage.

Thurstone, L. L. (1931) “The Indifference Function”, *Journal of Social Psychology* 2: 139–67.

Wallis, W. A., and M. Friedman (1942) “The Empirical Derivation of Indifference Functions”. In *Studies in Mathematical Economics and Econometrics*, edited by O. Lange, F. McIntyre, and T. O. Yntema. University of Chicago Press, 175–89.

Wiessner, P. (2009) “Experimental Games and Games of Life among the Ju/’hoan Bushmen”, *Current Anthropology* 50: 133-138.

Wilde, L. L. (1981) “On the Use of Laboratory Experiments in Economics,” in J. C. Pitt (ed.) *Philosophy in Economics*. Dordrecht: Reidel, pp. 137-48.

Zizzo, D. (2010) “Experimenter Demand Effects in Economic Experiments”, *Experimental Economics* 13: 75-98.

Zizzo, D. (2011) “Do Dictator Games Measure Altruism?”. In L. Bruni and S. Zamagni (eds.) *The Handbook on the Economics of Philanthropy, Reciprocity and Social Enterprise.* Edward Elgar, forthcoming.

Short biographies:

María Jiménez-Buedo is an Assistant Professor at the Department of Logic, History and Philosophy of Science at UNED in Madrid. She has worked in the fields of Political Economy, Science Policy, and Philosophy of the Social Sciences, with an emphasis on methodological issues.

Francesco Guala is a philosopher and experimental economist interested mainly in the foundations of social science. He is the author of *The Methodology of Experimental Economics* (Cambridge 2005), and of many articles in philosophy and social science journals. His next book, *Understanding Institutions*, will be published by Princeton University Press in 2016.