

# **Experiments in the Social Sciences: The relationship between External and Internal Validity**

**María Jiménez-Buedo\***

UNED. Departamento de Lógica, Historia y Filosofía de la Ciencia.

**Luis M. Miller**

Nuffield Centre for Experimental Social sciences. University of Oxford

The article identifies a latent debate in the recent literature on the role and worth of experiments in economics and other social sciences concerning the relationship between the external and the internal validity of experimental designs. Our work identifies two incompatible views regarding the relationship between internal and external validity of experiments. While in the methodological literature references to the idea that there is a trade-off between the internal and external validity of experiments abound, this view coexists with the position stating that internal validity is rather a prerequisite of external validity.

By identifying the contours of this implicit debate in the recent methodological literature around the use of experiments in the social sciences we call attention upon a series of insufficiently conceptualized issues regarding the central notions of internal and external validity and we question the standard view positing a trade-off between the two. This article stands against common associations of internal validity and external validity with the distinction between field and laboratory experiments and assesses critically the arguments that link the artificiality of experimental settings done in the laboratory with the purported trade-off between internal and external validity.

**Keywords:** internal validity, external validity, experiments, experimental economics.

---

\* Corresponding author: mariajimenezbuedo@gmail.com

## 1. Introduction

In the last two decades the debates around the worth of the experimental method in economics, and in general, in the social sciences, have been many, heated, and salient both for practitioners and methodologists. This is mainly a consequence of the consolidation of the experimental method as a valid tool for economic research which, as a side effect, has reopened the discussion about the benefits and drawbacks of laboratory experiments in the social sciences.<sup>†</sup>

Much of the methodological discussion around experiments in economics is framed in terms of the notions of internal and external validity, coined more than fifty years ago by Donald Campbell and his collaborators (Campbell and Stanley 1963; Cook and Campbell 1979; Shadish, Cook and Campbell, 2002). Internal and external validity appeal to us all as obvious requisites for the worth of an experiment. If an experiment is not internally valid, then, we cannot say that the treatment given in the experiment is the cause of the effect we observe. If an experiment is not externally valid, then its results cannot be said to hold outside of the experimental setting, and thus, even if internally valid, we cannot say anything relevant of the world. Quoting the classical definitions of Cook and Campbell (1979, 37), internal validity “refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause”, and external validity “refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times.”<sup>‡</sup>

---

<sup>†</sup> Examples of this emerging interest in the experimental method are three recent books (Guala 2005; Willer and Walker 2007; Webster and Sell 2007) and a recent issue of *The Journal of Economic Methodology* (2005).

<sup>‡</sup> Campbell and his collaborators have also invoked other types of validity regarding social science experiments, and their classification of validity types has evolved over time. Their most stable and recent list includes now four of them: *statistical conclusion validity*, *internal validity*, *construct validity* and *external validity* (Shadish, Cook and Campbell 2002). However, the most often invoked ones in both, experimental economics and theory-driven experiments in the social sciences, are those composing the internal-external validity dyad, to the puzzlement of some of the methodologists reflecting on the reception of the categories introduced by Campbell and his collaborators (for a brief discussion on the use of these categories in experiments in political science see, for example, Morton 2008)

Although most of the current arguments and disputes around the use of experiments in the social sciences refer to either type of validity, it is surprising that not much has been written systematically about the relationship between the internal and external validity of experiments. Following the Campbellian approach to validity stated above, the standard view among both methodologists and practitioners of experiments is that internal validity and external validity stand in a relationship best described as a *trade-off*: the more we ensure that the treatment is isolated from potential confounds in order to ensure that the observed effect is attributable to the treatment, the more unlikely it is that the experimental results can be eloquent of phenomena of the outside world, since typically, in the outside world, many factors interact in the production of events that we are interested in. References to this tension are common both in social psychology (Brehm, Kassin and Fein 1990, 45; Smith and Mackie 1999, 43) and economics texts (Guala 2002, 262; 2005, 144).

Although this seems to be the standard view regarding the relation between internal and external validity, it is not the only one. Within the same methodological and philosophical literature, we often find another idea that shares with that of the *trade-off* between internal and external validity the quality of having an *ipso facto* credibility appealing to commonsense: unless we ensure internal validity of an experiment, little, or rather nothing, can be said of the outside world. For some, thus, internal validity is in this way a *prerequisite* to external validity (Thye 2000, 1303; Lucas 2003, 248; Guala 2003, 1198; Hogarth 2005, 262).

How well do these two differing views on the relation between internal and external validity stand together? Can one simultaneously maintain that there is a trade-off between internal and external validity and that internal validity is a prerequisite for external validity? Although these two positions need not necessarily be contradictory, they do not stand in an easy relation to each other. It is therefore puzzling that no systematic attempt to address the implicit debate between these two views has arisen.

This article addresses the problem of the compatibility of these two ideas and analyzes critically the standing arguments about the conditions under which a trade-off between internal and external validity arises. Our contention is that the fact that this debate is still underdeveloped within the relevant literature shows that there is an array of questions that have yet to be more thoroughly conceptualized regarding the notions of internal and external validity and their current uses. Our argument stands against common associations of internal validity and external validity with the distinction

between field and laboratory experiments and assesses critically the arguments that link the artificiality of experimental settings done in the laboratory with the purported trade-off between internal and external validity.

The rest of the article is structured as follows. First, we review the arguments that have been put forward that posit a trade-off between internal and external validity followed by a review of the arguments that claim, instead, that internal validity is a prerequisite of external validity (section 2). We then discuss analytically the assumed tension or trade-off between internal and external validity and provide a series of criticisms to the standard view on why this trade-off may arise (section 3). In the light of this discussion, we then analyze a well-known example of a field experiment in economics to illustrate our argument and we contrast it with the arguments that the received views on the internal and external validity of lab versus field experiments would yield (section 4). In view of our analysis, we argue that there are no grounds to posit a general thesis about the way in which internal and external validity are related. Finally, in a concluding section, we draw some implications regarding the usefulness of the notions of internal and external validity as they are commonly conceptualized in the experimental economic literature.

## **2. The latent debate around the idea of a trade-off between internal and external validity**

The existence of a trade-off between internal and external validity constitutes a commonplace both in the experimental and in the methodological literature around experimental economics, and more broadly, in other disciplines where experiments are part of the common practice of scientists. One of the most influential books reflecting on the philosophical debates that surround experiments in economics asserts, for example that: “There is a trade-off between the internal validity of an experimental result (whether a given laboratory phenomenon or mechanism has been correctly identified) and its external validity (whether the results can be generalized from the laboratory to the outside world)” (Guala 2005, xi). Campbell himself referred in various occasions to this idea when reflecting on the notions of internal and external validity: “Both types of criteria are obviously important, even though they are frequently at odds

in that features increasing one may jeopardize the other” (Campbell and Stanley 1963). In other works we find references about a *tension* between both types of validity, as for example: “There is an *obvious* tension between the two. Where internal validity often requires abstraction and simplification to make the research more tractable, these concessions are made at the cost of decreasing external validity” (Schram 2005, 226, *emphasis added*).

An equally intuitive idea in the literature is that internal validity is actually a precondition of external validity. Hogarth, for example, has put it in the following terms: “internal validity is a necessary but not sufficient condition for external validity” (2005, 262). Guala himself, in a precedent work to his renowned book has affirmed that “problems of internal validity are chronologically and epistemically antecedent to problems of external validity” (Guala 2003, 1198), and similar arguments can be found in Lucas (2003, 248). Thye has put this idea in clear, intuitive terms “if there are doubts or questions about whether a relationship is real or spurious, then whether or not the finding applies to other settings is irrelevant” (2000, 1303).

Now, although these two ideas, i.e., that internal and external validity stand in a trade-off relationship with each other, and that internal validity is a prerequisite to external validity, are not necessarily *ipso facto* incompatible, it is at the very least not clear how to combine the two. In fact, we have not found simultaneous references to the two ideas within the same article, or explicit accounts about the ways or circumstances under which both are simultaneously defensible. What are the conditions under which one variable can have precedence over another, yet both stand in a trade-off relationship with each other?

Part of the difficulty of answering this question resides in the fact that the notion of a trade-off is *per se*, vague or elusive: what do we exactly mean when we say that two variables stand in a trade-off to each other? At a very basic, intuitive level, a trade-off between two variables must imply that the more we get of one, the less we get of the other. If we want to make the question of the trade-off between internal and external validity analytically tractable, we can take this general assertion about this trade-off to mean two differentiated things: A trade-off between internal and external validity can imply that either we can only obtain experiments that have low internal validity and high external validity and vice versa, or that in a given experimental design, internal validity can be increased at the expense of external validity and vice versa. We will disregard the first of these claims as implausible given that there seems little room for

arguing against the existence of paradigmatic experiments that are *both* internally and externally valid and we will focus on our efforts to disentangle the second of these claims, i.e., that in a given experimental design, external validity and internal validity can be exchanged for or traded-off against each other.

The relevant claims for affirming the existence of a trade-off between internal and external validity can thus be stated either as (1): in a given experimental setting, the design can be altered in order for the inferences from the experiment to have more internal validity at the expense of external validity or (2): in a given experimental setting, the design can be altered in order for the inferences from the experiment to have more external validity at the expense of internal validity. Once we put the trade-off between internal and external validity in these terms, what needs to be assessed therefore is whether either, both, or none of propositions (1) and (2) above is compatible with the idea stating that internal validity is a prerequisite of external validity.

An answer to this question would have implications: in order to simultaneously hold that internal validity is a necessary condition of external validity and that there is a trade-off between internal and external validity in the above-mentioned sense then this trade-off cannot be *symmetric*: if internal validity is a prerequisite of external validity then experimental designs may be altered in order for them to have more internal validity and less external validity, but not the opposite. Or at least this would be the conclusion if by asserting that internal validity is a prerequisite of external validity we meant that all the external validity of an experiment stemmed from its internal validity. But again, this idea seems rather difficult to grasp and has a certain degree of counterintuitiveness. If we accept that within an experimental design, internal validity is the source of external validity, how then, can one tamper with the experimental setting in order to gain internal validity by trading it for external validity? Once internal validity is conceived as the source of external validity, changes in the experimental design that increase internal validity should only help to enhance the external validity of an experiment. It seems therefore rather difficult to simultaneously hold the thesis that internal and external validity stand in a trade-off relationship with each other and that internal validity is a prerequisite for external validity: the debate is there and still unaddressed by the relevant literature. The fact that this incompatibility has not yet been subject to systematic scrutiny actually points to the existence of important lacunae in the recent methodological literature.

First, there seems to be an insufficient degree of conceptualization in the literature about the conditions under which internal and external validity of experiments are inversely related or at odds with each other. Second, and more sternly: we do not seem to know enough either about the concepts that yet shape a big part of the core of our methodological discussions on experiments, i.e., internal and external validity. This fact becomes patent once we acknowledge that there is actually no consensus about the right answers to an array of very basic questions about these concepts, some of the questions being: Is internal validity the minimum requisite to the relevance of an experiment? Is internal validity a quality that an experiment either has or not, or can experiments be more or less internally valid? Can an experiment, as some claim (Vissers et al. 2001; Kanazawa 1999) be relevant even if it has no external validity or makes no claims about the generalizations of causal claims outside laboratory conditions?

This article, by critically analyzing the standard view regarding the relationship between internal and external validity tries to call attention upon what we think is an insufficient degree of conceptualization of the internal/external validity dyad, which remains a central conceptual tenet in the recent experimental economics literature. In the following section, we spell out an analytical framework for the examination of the ways in which internal and external validity may be at odds with each other within social scientific experimental settings.

### **3. Is there a tension between internal and external validity? When?**

Let us recall the structure that characterizes the general logic of an experimental design in the social sciences. A very basic scheme of a perfectly controlled experiment is normally described in the following way (see table 1).<sup>§</sup> Ideally, the aim of an experiment is to isolate one single factor by comparing a treatment group (exposed to this factor) with a control group (not exposed). The experimenter tries to make sure that all other factors that might make these two groups different are kept constant. When and if indeed there is a true control over all other potential confounding factors possibly

---

<sup>§</sup> The summary is borrowed from Guala (2005).

influencing the variable of interest ( $Y$ ), a difference between  $Y_2$  and  $Y_1$ , (i.e.,  $Y_2 - Y_1 \neq 0$ ) is interpreted as the effect stemming from the (no longer putative, but accepted) cause.

**(Table 1 about here)**

This control over potential confounds is normally considered especially difficult in the social sciences due to the researcher's ignorance about all of the background factors potentially having an impact on the casual relation at hand (Shadish, Cook and Campbell 2002). Ensuring that the two groups differ in only one aspect is done, under perfect conditions, by direct control, and when this is not possible, then it is done by randomization and in that case we duly describe these designs as *randomized experiments*. Experiments in which units are not assigned to treatments randomly are, in Cook and Campbell's terminology, *quasi-experiments*; but if -as it is often customary- experiments are defined as studies in which an intervention is deliberately introduced to observe its effects (Shadish, Cook and Campbell 2002), then they too are, for short, referred to as experiments.

In theoretical terms, thus, according to this scheme, how can we conceptualize the supposed trade-off between internal and external validity? Suppose that a researcher is lucky enough in a given area of research as to have identified a cause, to have equally identified all of the other background factors potentially affecting the cause and to have managed to keep them constant, thus being able to attribute the difference between  $Y_2$  and  $Y_1$  as the effect of the treatment  $X$ . The results of the experiment in question would thus be internally valid. What about external validity? In what sense, or under what conditions would have she sacrificed external validity in such a setting, in order to reach internal validity?

In the methodological literature regarding experiments in the social sciences, the most common argument as to why external and internal validity stand in a trade-off relationship is one about the *artificiality* of experiments: because the experimental setting has been constructed by the experimenter, precisely, in order to ensure internal validity, then we cannot be sure that the causal mechanisms involved in the experiment hold outside the laboratory, and therefore, there are grounds to doubt, or at least, we cannot be certain that the phenomena identified under controlled circumstances does

hold in the outside world. Thus, the zeal with which experimenters ensure internally valid results goes counter to their capacity to extrapolate findings to what should be considered relevant (i.e., real world) conditions. Guala (2005, 144) has argued that the trade-off occurs because of shielding experimental system from random disturbances: “The more artificial the environment, the better for internal validity, the less artificial, the better for external purposes.” In the same vein, Schram refers to artificiality in terms of abstraction and simplification, and he argues that: “Where internal validity often requires abstraction and simplification to make the research more tractable, these concessions are made at the cost of decreasing external validity” (2005, 226).

In a nutshell, the standard argument is this: the very advantage that experiments bring about, i.e., an *artificially* controlled environment where putative causes can be isolated from other background factors so that effects can be soundly attributed to causes, makes the inferences from the experiment to the real world –our ultimate interest– difficult or problematic. Now, two questions may arise: is this idea sound? And even if the idea makes sense, is it best expressed by referring to it as a trade-off between external and internal validity?

One of the problems with this view linking artificiality, internal, and external validity is, again, its vagueness. For one thing, artificiality as an attribute of experiments is a rather more elusive concept than we might normally acknowledge. In this sense, Vissers et al. (2001) have argued that “artificiality” in defining the degree of intervention in experiments in the social sciences defies a precise definition. Their argument suggests that the degree of artificiality as a characteristic of an experimental design falls on a continuum (experimental settings ranging from *more* artificial to *less* artificial) that depends ultimately on the subjective view of *both* the experimenter and the experimental subject. Once we acknowledge that artificiality is hard to define, mapping it onto the field/laboratory distinction seems thus problematic, and assuming that laboratory experiments are necessarily more artificial than field ones is ungranted. This rather obvious point, needs perhaps to be stressed in view of how common the references in the literature as to the opposite. It suffices perhaps to exemplify it with a simple (and most unlikely) illustration. Suppose, for the sake of argument, that an experimenter selects two remote communities (equivalent in every relevant aspect) for which the institution of money is unknown and then introduces monetary units in one of them in order to measure how the existence of monetary exchanges will affect individual access to basic goods in that community. Contrast this situation with familiar

laboratory experiments in which subjects are asked to perform well-known tasks like completing simple math tests to check whether achievement rates depend on the introduction of monetary rewards associated to completion. Would we want to say that our hypothetical field experiment is less artificial than the laboratory one in virtue of it being performed in “real life” conditions?

At this point, it could be objected, nevertheless, that artificiality in a social science experiment may remain a crucial threat to the validity of the inferences that are made from it due to the fact of the existence of “investigator effects”, or what is also referred to as the “Hawthorne effect”, i.e., the systematic influence on subjects by the fact that they are knowingly being under scrutiny. Whether this effect is pondered and either accounted for or neutralized in each particular experimental setting may crucially depend on the skills and care with which each individual experiment is conducted, and may be circumvented if in field settings subjects are not unaware that they are being subject to an experiment or if in laboratory settings their performance is measured in ways that are opaque to the experimental subjects. Taking into account that this effect may in the end, and in different degrees, be an unavoidable part of any social interaction (as argued by Vissers et al. 2001), akin to the problem of reflexivity in the social sciences, it in any case seems to be unrelated to the trade-off between internal and external validity: if the Hawthorne effect is to interfere with the validity of experimental results, then it should be inimical to both the internal and external validity of the inferences that one may want to make from the experiment.

If artificiality is instead taken to mean something like manipulability, the conclusion is the opposite as in the case of “investigator effects”, yet equally unrelated to a trade-off between internal and external validity: Manipulability of the crucial variables involved in an experiment cannot be linked to an inverse relation between internal and external validity since it sets out the very possibility of experimentation, and is the base from which both the internal and external validity of the inferences from an experiment may be valid.

There is a further way in which artificiality could be conceptualized: if artificiality refers to the difficulty in inferring causality in the outside world from processes that are only found in experimental conditions \*\*, then it is best represented as

---

\*\* Peacock (2007, 9) highlights some standard procedures in experimental economics that make the laboratory situations not sufficiently “lifelike”, i.e., *artificial*. For instance, “situations in which players’

a threat to other types of validity of the experiment, namely, and foremost, to what Campbell and his collaborators referred to as construct validity, or the validity with which inferences can legitimately be made from the operationalizations in a particular study to the theoretical constructs on which those operationalizations are based, i.e., the “validity of inferences about the higher order constructs that represent sampling particulars” (Shadish, Cook and Campbell 2002, 65), or the validity of generalisations from operations to constructs (Caamaño Alegre 2009, 26). It is then important to note that this type of validity was first listed by Campbell and Stanley under the general heading of external validity and that Cook and Campbell (1963) later stated that it was intrinsically related to the problem of generalization (Cook and Campbell 1979, 38-39). Seen in this way, artificiality seems to be able to account for a kind of threat to the external validity of an experiment, and not for the tension between internal and external validity.

In sum, artificiality seems a vaguer notion that normally acknowledged, and depending on the more precise meaning it may take it may represent either different threats to the validity of inferences from the experiment, or none at all (as when linked to manipulability). It seems thus that none of the possible meanings of artificiality make it a suitable mechanism behind the supposed trade-off between the internal and external validity of experiments.<sup>††</sup>

The claim linking artificiality to the trade-off is very probably behind the widespread belief in the idea that field experiments are strong in external validity and weak in internal validity and that, inversely, laboratory experiments are strong in internal validity and weak in external validity (List and Levitt 2004; Lusk and Norwood 2006). Again, this type of reasoning seems based on intuitions rather than on analytical

---

anonymity is upheld [...]; situations in which encounters are ‘one-shot’ in nature [...]; situations in which, by experimental design, players can never be matched against the same player more than once.”

<sup>††</sup> Artificiality is however not the only mechanism that has been purported in order to account for the trade-off between internal and external validity. For example, Shadish, Cook and Campbell (2002) have linked this trade-off to the use of different methods for ensuring the reliability of inferences from the experiments, and so they claim: “the best-known example is the decision to use randomized experiments, which often helps internal validity but hampers external validity” (p. 34). However, the factors that they claim as linking randomization to a trade-off between internal and external validity are entirely pragmatic in the least thrilling sense of the term, and so they state (p. 96): “in a world of limited resources, researchers always make tradeoffs among validity types in any single study. [...] (R)andom assignment [of treatments] can help greatly in improving internal validity, but the organizations willing to tolerate this are probably less representative than organizations willing to tolerate passive measurement, so external validity may be compromised”.

scrutiny. In terms of the basic structure of experiments depicted in the previous section, this association between field experiments having less internal validity and laboratory experiments having less external validity, cannot be argued for, for the theoretical structure of both experiments, described above, is the same for field and laboratory experiments. The idea of field experiments being *intrinsically* less amenable to *control* and laboratory experiments being *intrinsically* less amenable to *extrapolation* is, as soon as examined critically, ungranted. It is only logical to think that the capacities to control in the experiment or to extrapolate to non experimental contexts rather depend, crucially, on the established background knowledge that the experimental scientist possesses about potential confounds that intervene in the causal process under scrutiny via the experiment. The availability of factors warranting that either *matching* or *randomization* may be exhaustive for the control of variation in only *X* depend thus on the intrinsic characteristics of the causal process that is being studied and in our previous knowledge of it, rather than on whether the experiment is done in a lab or in a field setting.

At this point one might object that the association between the lab providing high internal validity and the field high external validity makes reference to just a *tendency*, and that it is not to be interpreted as an analytical statement but as an empirical regularity. But even then: would those siding with this less ambitious claim be ready to assert that, generally speaking, laboratory experiments involve claims and inferences about causal processes about which we have better background knowledge of confounds than field experiments? This more modest claim seems equally hard to sustain: surely, the availability of well-accepted knowledge about potential confounds depends on factors like previous familiarity with the causal hypothesis being studied or tested, on the maturity of the research program to which the causal hypothesis pertains and on various other general aspects that normally determine what we consider to be either well-established or tentative knowledge. This, in turn, has little relation to whether experiments are performed in the laboratory or in the field.

In order to further illustrate the problematic nature of the thesis positing a trade-off between internal and external validity, and related claims in terms of the distinction between field and laboratory experiments we now examine the case of well known, much cited field experiment by Uri Gneezy and Aldo Rustichini (2000).

#### **4. A well-known field experiment: Gneezy and Rustichini**

## on the *deterrence hypothesis*

Gneezy and Rustichini conducted, at the end of the 1990s, a field experiment involving 10 kindergarten schools in the city of Haifa, in Israel. The experiment was meant to test the deterrence hypothesis, widespread in legal studies and the basis of some psychological work on behaviour modification: the introduction of a penalty, *ceteris paribus*, reduces the occurrence of the behaviour subject to the fine. Day-care centers normally face the common problem of parents arriving late to collect their children: the experiment testing the deterrence hypothesis consisted in introducing a fine in six of the ten day-care centers. A flat rate fine was imposed on those parents that arrived ten or more minutes late. The other four centers, where everything was left unchanged, served as a control group. The treatment (a monetary sanction) was assigned to six of the day-care centers that were identical, in every relevant respect,<sup>‡‡</sup> to the other day-care centers that served as the control group. Parents in the treatment group were informed by the managers of the day-care centres of the introduction of the fine but were unaware that they were being the subjects in an experiment. The number of parents coming late was then measured and found significantly higher in the treated population: the study showed that in those centres where the fine was introduced there was an increase in the number of parents coming late, thus contradicting the deterrence hypothesis. The authors of the article favour an explanation in terms of incomplete contracts and information: in the absence of a fine, parents cannot be certain about the consequences of misbehaviour (like arriving late to pick up their children, imposing a burden on the teacher that has to stay longer to take care of the child) and so tend to comply to the rule of arriving in time for fear of the unspecified consequence. Once a fine is imposed, they can be certain of the perceived cost of their behaviour on the part of the managers of the school, and so some parents that were restraining themselves from arriving late will now do so, knowing that they will be fined by the specified amount. The fine thus serves as a price that conveys information on the cost of their behaviour, and arriving late becomes a “commodity”. This serves to explain a second, puzzling, finding of the study: the experiment lasted twenty weeks. Fines were introduced in the treated schools on the fifth week and lifted on week number

---

<sup>‡‡</sup> The authors state this explicitly: “All of these centers [both those under the treatment and those under the control conditions] are located in the same part of the town, and there is no important difference among them” (p. 4).

seventeen. The increase in the number of parents arriving late after school was nevertheless maintained even after the fine was lifted, a fact explained by the authors by the fact that once coming late became a commodity with a well-known price, then it remained one even in the absence of a fee, or in their words: “once a commodity, always a commodity” (p. 14).

In terms of the internal and external validity of the inferences drawn from this experiment, the textbook or standard approach on the matter would suggest that given that their experiment was carried out in the field, the type of control and manipulation that is achievable in a laboratory and that helps ensuring internal validity is less likely to have been attained. On the other hand, though, because the experiment is carried out at the actual premises where the researchers would ideally want to test their deterrence hypothesis, the question on whether the relationship found between the treatment and the effect *outside* of the experimental conditions does not even pose itself, because the experimental and the real conditions are the same. The received view about external validity in the experimental economics literature would therefore suggest that the fact that this experiment is done in real conditions and with real fines can make us fairly confident that the relationship found in the experiment can be generalizable to other similar, parallel situations.<sup>§§</sup> The question this article poses is though: is this standard view on the usual properties of field experiments in terms of internal and external validity useful, enlightening, or true in some sense of this experiment in particular and of field experiments in general?

Upon examination, we can say that it is indeed the case that Gneezy and Rustichini’s study faces at least a clear threat to validity, that is (at least in written) overlooked by the authors. This threat is typically operative in social science experiments and is identified by what Cook and Campbell coined “treatment diffusion”<sup>\*\*\*</sup>. The fact that the treatment is conducted in 10 day-care centres in the same city opens the possibility of parents being aware of the fact that they are taking part of an experiment. This in turn can interfere with the response to the treatment. Actually, G&R’s results, where the behaviour observed consists in many of the parents arriving

---

<sup>§§</sup> Note again, in terms of the implicit debate in the literature that this article has underlined, that this stereotyped view on the validity of field versus laboratory distinction is in turn problematic with respect to the view that internal validity is a prerequisite of external validity: field experiments facing serious threats to internal validity should have external validity problems.

<sup>\*\*\*</sup> Treatment diffusion is classified as a typical threat to internal validity in Campbell’s earlier works (Cook and Campbell 1979), though is listed among construct validity threats in his later writings (Shadish, Cook and Campbell 2002)

late once a fine is introduced could be congruent with the behaviour of offended parents that, having found out that they are being *treated* (either in the experimental sense, or quite simply, in the more pedestrian sense of the word) differently, decide to rebel against the treatment by arriving later and not earlier to school, as a form of resistance of protest.

Our point of interest here though, is not to signal this as a viable alternative explanation for the behaviour observed to the one provided by G&R, since, it could be argued, for example, that the behaviour observed after the fine is lifted (where parents that had been subjected to the fine continue to arrive later than before the fine was introduced) seems less congruent with the treatment diffusion hypothesis. Our purpose in pointing this out is, instead, to underline the fact that the design presented and defended by G&R is, indeed, subject to at least some validity problems that could be attributed to the fact that the experiment is done in the field. The field context of the experiment makes isolation of subjects impossible in practice and therefore the possibility exists for an unwanted diffusion of the treatment amongst members of treated and control groups. In the remaining of our discussion we will however ignore this potential threat to the validity of the results. First, as we signalled just above, had this threat been effective it would be difficult to explain that parents continued to arrive late after the fine was lifted. Second, at least in theoretical terms this threat is only half-attributable to the experiment being performed in the field. Had Gneezy and Rustichini more resources and, were it possible to draw enough parallels in terms of equivalence (in relevant aspects to the experiment) between cities, the experiment could have been performed simultaneously in more than one city, in this way allowing the experimenters to choose only one school per town. It seems thus that this particular threat to the validity of inferences from the experimental results stems more from feasibility constraints related to the resources devoted to research than from intrinsic properties of the field.

A more interesting characteristic of the experiment is, however, inseparable from the fact that it is a field experiment, and according to Gneezy and Rustichini it stems, precisely, from the fact that there is no full control over all the factors other than the treatment (the fine) affecting the variable under study (behaviour over punctuality in picking up one's own children at the day-care centre). So, as G&R say themselves, "we argue that penalties are usually introduced into an incomplete contract, social or private. They may change the information that agents have, and therefore the effect on

behaviour may be opposite of that expected. If this is true, the deterrence hypothesis loses its predictive strength, since the clause "everything else is left unchanged" might be hard to satisfy" or, in their conclusions: "the effect of a change in a clause of the contract may produce effects different from what might be expected from the assumption that "everything else is left unchanged"". In other words, G&R argue that because the introduction of a fine modifies other factors affecting the variable of interest (normally labelled as confounding factors), the change in behaviour is different that the theory predicts, for the theory makes, precisely, predictions that rest on a *ceteris paribus* clause that does not take place in real life conditions. The aim of G&R's experimental design is actually that of contrasting the behavioural predictions of the deterrence hypothesis under isolated conditions, typically found in a theoretical model, from the predictions that would stem from real life conditions, and in particular, from the incompleteness of contracts that often takes place in real life conditions. In this sense, and in common with a vast proportion of the experiments in behavioural economics, G&R's work tries to contrast the agents' behaviour under the treatment and control with a *theoretical* prediction stemming from a *theoretical* model. The predictions stemming from the theoretical model can in turn be interpreted in terms of the same logical structure that is used to understand the general logic of experimental design (see figure 1 below), where the comparative statics of the model provide an account of the effect of changes in one variable with respect to another.<sup>†††</sup>

G&R's work can thus be interpreted as an attempt at showing that in the presence of a particular confound (incompleteness of contracts), the relationship between an externally imposed material cost related to an action and the frequency of the action do not relate in the same way as they would do in the absence of that element (or, as represented in Table 2 below:  $[(Y_2 - Y_1) \neq (Y_4 - Y_3)]$ ). The findings in each case being that  $(Y_2 - Y_1) > 0$  and  $(Y_4 - Y_3) < 0$ ).

**(Table 2 about here)**

---

<sup>†††</sup> The difference though between the theoretical model and the experiment resides in the fact that the model a priori *ensures* the isolation that the experiment can only *aim* at providing. For an extended account of the parallelisms and differences between models and experiments in economics see Morgan (2005), Maki (2005) and Alexandrova (2006).

Put in this manner, we can see how G&R's way to point at the relevance of their results rests on having identified experimentally a variable or a background condition that mediates the relationship between  $X$  and  $Y$  in a way that standard theoretical models had yet not captured. We can in principle assume that were it possible to recreate the relevant variables and background condition(s) in the laboratory (i.e., introducing fines, arriving late, the incompleteness of contracts), the authors would probably (or at least they would have had no reason not to) have carried out a laboratory rather than a field experiment. Given that the variable  $Y$  (arriving late to pick up one's children) and possibly the relevant background condition (incompleteness of contracts) are not easily manipulable in the lab, then the choice of a field experiment seems apt to G&R's purposes. The example thus shows that the standard view on internal validity, linking it to the laboratory experiments, seems inadequate in this case: given the inferences that the authors aim at with their experiment, and given that manipulability of the crucial variables can only be reproduced in the field, a laboratory would have not ensured the internal validity of the causal inferences that the authors make, but to the contrary. Upon examination of the kind of causal inferences claimed from the results, one is to suppose that the authors' choice for a field rather than a laboratory experiment is pragmatic and based on the need to introduce variables that may be difficult to manipulate in a laboratory, rather than to avoid artificiality that according to the standard view would secure internal validity at the cost of external validity of results.

As we pointed out already, and as regarding external validity, the standard view on the literature could lead us to assume that G&R's experiment should in principle be little problematic: the experimenters should have no problem in justifying the proposed conclusions outside experimental conditions because the experimental and outside conditions coincide. However, about external validity, though, the experiment's authors themselves, and, despite what is commonly assumed of field experiments, do not make but very modest claims on generalizability of results: G&R actually make sure they mention at several points in their work that their findings are not necessarily generalizable to other situations and that their results may be dependent on some of the idiosyncrasies of their experimental setting, like for example, the particular size of the fine that they have chosen for their study. Contra the commonly held view, the authors seem to endorse the view that it is the concreteness of the experimental setting and not its artificiality which makes the results less rather than more generalizable. Their

position can be spelled in the following way: since the experimental results are the product of a concrete setting we cannot be certain (even if we take as valid G&R's hypothesis stating that the relevant intermediating variable between the fine and the observed behaviour is the incompleteness of contracts) about which of any of the other elements present in the experimental situation may have helped to trigger the result. In this way, we must at least contemplate the possibility that the incompleteness of contracts can have the purported effect only in the presence of yet another intervening factor. In view of this, one could take, as candidates for latent mediating variables, an arbitrarily long list of factors that are present in G&R's concrete experimental setting: we could thus conceive of the fact that the observed behaviour is due to the introduction of a fine, but that it will only take place when the experimental agents are under stress (as in late-arrival parents), or that those results can only to be found in Meridional countries where there is a particular relationship towards social norms, or only when the reputation of agents vis-à-vis a mildly hierarchical (i.e., with limited yet unknown rule-enforcing capacities) actor is involved (as in kindergarten directors and teachers), etc. Note that this latter argument seems to point in a direction that would lead us to endorse the view that internal validity is actually a prerequisite for external validity: because we cannot be certain of having isolated all the potential variables intervening in the process under study, generalizations from the experiment seem difficult or ungranted. Yet, we should recall here that isolation of a certain sort seems to run counter to manipulability in this particular case, and therefore also against internal validity.

G&R's example thus shows that the standard arguments accounting for the association between internal validity and laboratory, and external validity and field experiments rest on an unsatisfactory conceptualization but, and going back to our question of concern: how are, in this experiment, internal and external validity related? Did, in this particular field experiment, external validity come at the expense of internal validity or vice versa?

In order to overcome the limitations regarding the generalizability of the experiment, or its external validity, any methodologically concerned reader would prescribe that G&R replicate their findings to similar, parallel, situations (e.g., that they rerun the experiment in other countries or in settings that differ from kindergarten centres yet relate to it in relevant aspects, like the workplace or to the assisting to meetings in voluntary associations). In turn, and in order to overcome the limitations of the experiment in terms of internal validity, the most obvious prescription seems to be

shockingly similar to the one associated to improving external validity: the best and more straightforward way to reduce the potential for alternative explanations to the observed behaviour or to isolate from confounding effects would also be to reproduce the experiment in different contexts.

If reproducing the experiment in other settings is what one would prescribe in order to overcome potential threats to both internal and external validity, then it is difficult to argue that internal and external validity stand in a trade-off relationship with each other, since such trade-off would require that one could conceive of changes in the experimental setting that would have allowed for an increase in internal validity at the expense external validity or vice versa.

## **5. Conclusion**

Scientists characteristically carry out experiments in their unremitting quest for regularities and robust new findings. When threats to the validity of results appear and are realized, attempts to control for these threats lead them or fellow experimenters to introduce changes in experimental designs. In experimental economics and theory-driven experiments in other social sciences, it is common to find mentions to threats to the internal validity relating to the concrete form of the incentives introduced in the setting and/or the interpretations that experimental subjects make of them. In terms of the generalizability of the results of these experiments, threats to external validity typically take the form of the unwarranted character of inferences from observed behaviour onto different groups of subjects or different contexts. In part, these are commonly found sets of problems because typically, the theories that are being tested by experiments do not address neither of these sets of issues explicitly. Yet, the inferential problems pertaining to both internal and external validity of experiments are very often characterized by the same logical structure, and the avoidance of threats to either type of validity often involves the same recommendation: the replication of the experiments under slight variations in order to account for potential confounds that may be partially responsible for the observed behavioural results.

The aim of this article has been to address a latent debate in experimental social sciences literature on the relationship between internal and external validity of experiments. We have adopted a critical stance to the standard position on this debate

by trying to show that problems of either external or internal validity do not crucially depend on the artificiality of experimental settings nor on the laboratory-field distinction between experiments. Our view is that that threats to internal or external validity depend on the particularities of the design and on problems with the operationalization of crucial variables, yet there seems to be no grounds posit a general trade-off between the internal and external validity of experiments.

The idea of a trade-off or tension between internal and external validity seems, upon analysis, far less useful than its intuitive attractiveness may lead us to think at first sight. The problem, though, may not reside solely in how this tension has been (little) characterized but in other important issues regarding the current definitions and uses of the concepts of internal and external validity. Methodological discussion and philosophical debate around these concepts, widely employed by social science practitioners and experimenters, is still needed.

## REFERENCES

Alexandrova, A. 2006. Connecting Economic Models to the Real World: Game Theory and the FCC Spectrum Auctions. *Philosophy of the Social Sciences* 36: 173-192.

Brehm, S. S., S. M. Kassin and S. Fein. 1990. *Social Psychology*. Boston: [Houghton Mifflin](#).

Caamaño-Alegre, M. 2009. Experimental Validity and Pragmatic Models in Empirical Science. *International Studies in the Philosophy of Science* 23: 19-45.

Calder, B. J., L. W. Philips and A. M. TYBOUT 1982. The Concept of External Validity. *Journal of Consumer Research* 9: 240-244.

CAMPBELL, D. T. and J. C. STANLEY (1963), *Experimental and Quasi-Experimental Designs for Research*, Chicago, Rand McNally and Company.

Cook, T. D. and D. T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.

Gneezy, U., and A. Rustichini 2000. A Fine is a Price. *Journal of Legal Studies* XXIX: 1-18.

Guala, F. 2002. On the scope of experiments in economics: comments on Siakantaris. *Cambridge Journal of Economics* 26: 261-267.

----- . 2003. Experimental Localism and External Validity. *Philosophy of Science* 70, 1195-1205.

----- . 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Harrison, G. and J. List. 2004. Field Experiments. *Journal of Economic Literature* XLII: 1009-1055.

Hogarth, R. B. 2005. The challenge of representativeness design in psychology and economics. *Journal of Economic Methodology* 12: 253-263.

Kanazawa, S. 1999. Using Laboratory Experiments to Test Theories of Corporate Behavior. *Rationality and Society* 11: 443-61.

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

Maki, U. 2005. Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology* 12: 303-315.

Morgan, M. 2005. Models are experiments, experiments are models. *Journal of Economic Methodology* 12: 317-329.

Morton, R. and K. Williams. 2009. From Nature to the lab: Experimental Political Science and the Study of Causality. New York University: Mimeo.

Peacock, M. S. 2007. The Conceptual Construction of Altruism: Ernst Fehr's Experimental Conduct to Human Conduct. *Philosophy of the Social Sciences* 37: 3-23.

Schram, A. 2005. Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology* 12: 225-237.

Smith, E. R. and D. M. MACKIE. 1999. *Social Psychology*. Philadelphia: Psychology Press.

Thye, S. R. 2000. Reliability in Experimental Sociology. *Social Forces* 78: 1277-1309.

Vissers, G., G. Heijne, V. Peters, and J. Geurts. 2001. The validity of laboratory research in social and behavioural science. *Quality and Quantity* 35: 129-145.

Webster, M. and J. Sell. 2007. *Laboratory Experiments in the Social Sciences*. Oxford: Academic Press/Elsevier.

Willer, D. and H. A. Walker 2007. *Building Experiments. Testing Social Theory*, Stanford: Stanford University Press.

†††

**Table 1: General logic of an experimental design**

	Treatment (Putative cause)	Putative effect	Other factors
Experimental group	X	Y <sub>1</sub>	Constant
Control group	0	Y <sub>2</sub>	Constant

\*\*\*\*\*

**Table 2: The contrast between the predictions of a baseline theoretical model and G&R's experimental results.**

**Comparative statics in theoretical models**

Complete Contract	Variable X (fine)	Variable Y (lateness)	Other factors
Treatment	X <sub>fine</sub>	Y <sub>1</sub>	Constant
Control	0	Y <sub>2</sub>	Constant

**G&R field experiment**

Incomplete contract	Treatment (Putative cause: fine)	Putative effect (lateness)	Other factors
Treatment	X <sub>fine</sub>	Y <sub>3</sub>	Constant
Control	0	Y <sub>4</sub>	Constant