# WHY A TRADE-OFF? THE RELATIONSHIP BETWEEN THE EXTERNAL AND INTERNAL VALIDITY OF EXPERIMENTS*

**María Jiménez-Buedo**
**(mjbuedo@fsof.uned.es)**
**Departamento de Lógica, Historia y Filosofía de la Ciencia**
**UNED. Madrid**

**Luis M. Miller**
**Center for Experimental Social Sciences.**
**Nuffield College. University of Oxford**

* Earlier versions of this paper were presented at the 2008 International Network for Economic Methodology Conference in Madrid and the 2009 Meeting of the Society for the Philosophy of Science in Practice, in Minneapolis, under the title: "*Experiments in the Social Sciences: The relationship between External and Internal Validity*".
A final version is the paper is published in ***THEORIA.*** *An International Journal for Theory, History and Foundations of Science.* Vol 25, No 3 (2010). **The full text can be downloaded (Open Access Journal) at:**
http://www.ehu.es/ojs/index.php/THEORIA/article/view/779/700

## Abstract:

Much of the methodological discussion around experiments in economics and other social sciences is framed in terms of the notions of internal and external validity. The standard view is that internal validity and external validity stand in a relationship best described as a *trade-off*. However, it is also commonly held that internal validity is a *prerequisite* to external validity. This article addresses the problem of the compatibility of these two ideas and analyzes critically the standard arguments about the conditions under which a trade-off between internal and external validity arises. 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. We conclude that the idea of a trade-off or tension between internal and external validity seems, upon analysis, far less cogent than its intuitive attractiveness may lead us to think at first sight.

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.[1]

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 use its results to 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."[2] With nuances and slight (and unsystematic) variations attributable to the likings of particular authors, these definitions continue to be employed and widely quoted in the methodological literature on experiments in the social sciences.

[1] 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 special issue of *The Journal of Economic Methodology* (2005).

[2] 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 and Williams 2009).

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 the 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 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 present a review of common arguments positing 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 1). We then discuss analytically the purported tension or trade-off between internal and external validity and provide a series of criticisms to the standard arguments as to why this tension may arise, namely, via the artificiality of the experimental setting (section 2). In the light of this discussion, we then analyze a well-known example of a field experiment in economics to illustrate how standard arguments about the relationship between internal and external validity fail when applied to concrete cases. In particular, we argue against the received views on the internal and external validity of lab versus field experiments (section 3). 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 and we argue against the idea that there is a trade-off between the two. Finally, in a concluding section, we draw some implications regarding some of the confusions surrounding the notions of internal and external validity as they are commonly conceptualized in the experimental economic literature (section 4).

## 1. 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 book has affirmed that "problems of internal validity are chronologically and epistemically antecedent to problems of external validity" (Guala 2003: 1198), and a similar argument 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. A question comes to mind to anyone trying to connect these two ideas: what would be the conditions under which internal validity can have precedence over external validity, 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 mean exactly 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 a general assertion about this trade-off to mean two differentiated things: Either we mean that 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. In what follows we will focus our efforts on the second of these two claims. On the one hand, the first claim would deny, quite implausibly, the existence of experiments that are *both* internally and externally valid, which in the face of what we know about paradigmatic experiments in science would seem too radical a

view. On the other hand, it is 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, that finds more resonance in the methodological literature that mentions the tension between both types of validity.

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 claim 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 (so (2) above, would not seem tenable) . Or at least this would be the conclusion if by asserting that internal validity is a prerequisite of external validity we meant that the internal validity of an experiment is a necessary condition to that experiment having any external validity at all. But again, this idea seems rather difficult to grasp and is itself problematic. If we accept that an experiment cannot have external validity if it does not have internal 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 a prerequisite to external validity, changes in the experimental design that increase internal validity should either leave external validity unaltered or at best, should help to enhance it (and making also (1) above unsustainable). After all, then, it seems 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 insufficiently addressed 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 experimental 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 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.

## 2. 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. The basic scheme of a perfectly controlled experiment is normally described in the following way (see table 1).[3] 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 influencing the variable of interest ($Y$), a significant difference between $Y_2$ and $Y_1$, (i.e.,$Y_2$-$Y_1 \neq 0$) is

[3] The summary is borrowed from Guala (2005).

legitimately interpreted as the effect stemming from the (no longer putative, but accepted) cause.

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

| | Treatment (Putative cause) | Putative effect | Other factors |
|---|---|---|---|
| Experimental group | X | $Y_1$ | Constant |
| Control group | 0 | $Y_2$ | Constant |

This control over potential confounds is sometimes considered to be especially difficult in the social sciences, due to the researcher's ignorance about all of the background factors potentially having an impact on the causal 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, where 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, and 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 causal relation 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. Cartwright, for example, has stated that "(I)t is a well–known methodological truism that in almost all cases there will be a trade off between internal validity and external validity (…) The usual complaint here is about the *artificiality* of the circumstances required to secure internal validity (.)" (2007: 220; *emphasis added*). Guala (2005: 144; *emphasis added*) has argued that the trade-off occurs due to the shielding of the experimental system from extraneous disturbances: "The more *artificial* the environment, the better for internal validity, the less *artificial*, the better for external purposes". 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, some questions may arise: What are we really referring to when we talk about the *artificiality* of the experimental setting? And even if we agreed on what artificiality means, is it really responsible of a trade-off between external and internal validity? In our view, the main problem with the view linking artificiality, internal, and external validity seems to lie in its vagueness: as an attribute of experiments, the notion of artificiality is a rather more elusive concept than we normally acknowledge.

Schram (2005), for example, while placing artificiality at the core of his analysis of internal and external validity, does not characterize the notion in a precise manner. In his work, and in a way that is representative of other accounts of artificiality, the notion gets defined only *negatively*, mainly as the flip side of external validity, and as related to a somewhat hazy notion of the mirroring the real world in the lab. He mentions that experimental studies may be accused of artificiality when they "[do] not *reflect* the 'real world' " (p. 3, *emphasis added*) and adds that "[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". This lack of a

thorough spelling of the notion of artificiality is representative of the methodological literature regarding experiments in economics and the social sciences in general.

In this sense, Vissers et al. (2001) have addressed this difficulty by underlining the fact that taking "artificiality" as an attribute of the *degree of intervention* in social science experiments necessarily defies a precise definition. First, in their view, artificiality as a characteristic of an experimental design is a relative notion, and falls on a continuum (particular characteristics of experimental settings ranging from *more* artificial to *less* artificial) but second, and more importantly, Vissers and colleagues argue that artificiality may ultimately depend on the *subjective* view of *both* the experimenter and the experimental subject, and that there is not an a priori measure of which traits of a given experimental design will be perceived as more artificial by the subjects under a particular treatment. As a consequence, they conclude that criticisms about a *general* artificiality of the lab miss the point.

As Vissers et al. (2001) suggest, there is yet another, perhaps more viable candidate for defining artificiality of the experiment in the social sciences univocally: perhaps, what is common and intrinsically artificial of social sciences experiments, and may constitute a crucial threat to the validity of the inferences that are made from them is what we normally call "investigator effects", or what is also referred to as the "Hawthorne effect", i.e., the systematic influence on subjects of the fact of knowingly being under scrutiny. Against this idea, however, we would argue that 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 aware 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. Moreover, this effect may in the end, and in different degrees, be an unavoidable part of *any* social interaction and can thus show not only in experimental data but also in observational one (as argued by Vissers et al. 2001). The Hawthorne´s effect is thus in some respects akin to the problem of reflexivity in the social sciences, and not restricted to the experimental method. In any case, it seems a poor candidate to constitute the mechanism behind the alleged trade-off between internal and external validity: if the Hawthorne effect is the kind of artificiality that 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.

Finally, and especially in experimental and behavioural economics, artificiality is directed, as a criticism, towards typical, particular experimental procedures. Peacock (2007: 9) for example, highlights some standard procedures in experimental economics that make the laboratory situations not sufficiently "lifelike", or artificial. For instance, "situations in which players' 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...". In these cases, an argument about whether artificiality can account for a trade-off between internal and external validity would need to be made about each particular trait, rather than generally about experiments.

Take, for example, the anonymity of players' clause in most versions of the better-known experimental games. The argument of those defending that artificiality accounts for the existence of a trade-off between external and internal validity would probably be spelt out in the following terms: though anonymity ensures the internal validity of the experiment, it hampers its external validity. Predictions stemming from the experimental results (where players do not know each other) onto the real world (where agents may be tied by social links) will not be possible and so the anonymity clause hinders the external validity of the experiment. The problem with this view is that it too quickly makes a number of suppositions about the experimenter's assumptions and goals that are not necessarily correct, or that should at least be examined case by case.

In particular, this view seems to ignore that the aim of most experiments is not to make precise, exact predictions about concrete patterns of behaviour in definite situations, but to isolate and identify underlying causal tendencies that are then assumed to hold across different conditions, despite not being always separately observable. Going back to our example regarding the anonymity clause in many experimental treatments, the aim behind the clause can normally correspond to two different logics: either the experimenter's goal is to test the relation between the treatment and the variable of interest under anonymity (in which case no argument about the inferential problems associated to the artificiality of anonymity could be made), or else the experimenter's goal is to test the relation between the treatment and the variable of interest across or *independently* of the degree of acquaintance between subjects. In this latter case, and as we have previously pointed out with regard to the basic structure of experimentation, more than one strategy is available to the researcher. First, the

experimenter can decide to match the subjects according to their levels of acquaintance. Secondly, she can choose to randomize her experimental subjects across their degree of acquaintance. Finally, she can fix this variable at a constant level (for example, at the level of anonymity) and perform the experiment, which, given the standard number of experimental subjects normally available to researchers, seems to be a likely viable alternative in many cases.

Depending on the degree of background knowledge of the researcher about the interaction between the variables of interest and the variable describing the degree of familiarity between agents, the strategy of fixing the level of a background variable can indeed potentially pose difficulties if the aim of the researcher is to make concrete predictions about the way in which the variables of interest interact in a given, concrete, "real life" target situation. As we have just pointed out, and in behavioural economics, making such predictions is rarely the purpose of any particular experiment[4]. But more importantly, and going back to our main concern: if we are to maintain that the anonymity clause gives rise to a trade-off between internal and external validity, then the mechanism behind it relates to the *fixing* of the value pertaining to its related variable (degree of acquaintance) at a given level, and not, to the artificiality of the trait. Therefore, once more: understood in this sense, artificiality remains a poor candidate for the mechanism responsible of the supposed trade-off between internal and external validity.

There is yet a further way in which artificiality could be conceptualized and it is often so portrayed in common criticisms of the results of behavioural economics. Frequently, the empirical results of theory-driven experiments in economics and other social sciences are accused by external critics of failing to capture the theoretical entities that they are after (e.g., trust, reciprocity, etc.) in the concrete, *artificial*, operationalizations of those entities that are chosen by the experimenter (e.g., contribution of one individual's monetary tokens to a common pool of resources, willingness to share with fellow participants a given amount of money provided by the experimenter, etc). This type of artificiality, though unavoidable in many instances, may

[4] Indeed, most experimental science aims at identifying causal tendencies that hold across many situations, yet the issue is how we can know whether the goal has been achieved. To say that causal tendencies are "assumed" to hold is unsatisfactory for an empiricist. The view defended by philosophers like Cartwright and Guala is that the only way to know whether we have identified a genuine tendency is to perform several experiments across different initial conditions - i.e. to check whether the tendency is at work in several concrete specific settings. So in their view appealing to tendencies is not a solution, but a way to describe a problem to be solved case-by-case. An anonymous referee made this point.

indeed represent a problem for external validity, but would actually be doing little service to internal validity. Indeed, this type of artificiality may actually be best represented as 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 some of the threats to the external validity of an experiment, and not for the tension between internal and external validity.

A further point is in order regarding common misconceptions around the distinction between internal and external validity and the role of artificiality in the supposed trade-off. Once it is acknowledged that artificiality may after all be difficult to define (in the sense in which there might not always be a set of characteristics of experimental treatments that can be a priori and objectively be deemed artificial), mapping it univocally onto the field/laboratory distinction seems problematic, and thus, assuming that laboratory experiments are necessarily more artificial than field ones is ungranted. This rather obvious point needs perhaps to be stressed since references in the literature abound on how laboratory experiments have *per se* more internal validity but less external validity because they are more artificial and how the reverse is true for field experiments (Levitt and List 2007; Lusk and Norwood 2006). Whether or not this could, as a matter of fact, be the case, it clearly is not a logical necessity, unless we take *artificial* to mean anything that is performed in a laboratory. It suffices perhaps to exemplify the idea with a simple (and imaginary, or 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? As mentioned above, attributing artificiality to an experimental intervention must depend too on the experimental subjects' own considerations, and should not, as commonly assumed, be necessarily equated with any set of typical settings found in the laboratory relative to any other field setting.

In sum, artificiality seems a vaguer notion than normally acknowledged, yet none of its possible meanings make it a suitable mechanism behind the supposed trade-off between the internal and external validity of experiments[5]. Instead, and very often, what seems to be at work is the conflation of two distinct phenomena: while on the one hand it is hard to use the results of a given experiment to make exact predictions about a concrete real phenomenon in the world, this does not analytically amount to the existence of a trade-off between the internal and external validity of experiments. After all, the difficulty of making exact predictions about the world will not be eased (if anything, it may be intensified) by making concessions about the zeal with which we try to ensure that our experimental causal inferences are not spurious.

As briefly mentioned above, 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. Again, references to this idea very often seem to be based on general intuitions rather than on thorough analytical 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,

[5]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".

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 standard mention to the association between the lab providing a higher internal validity and the field yielding a higher external validity makes reference to just a *tendency*, and that this association is not to be interpreted as an analytical claim but as an empirical regularity. But even then: would those siding with this less ambitious statement be ready to assert that, generally speaking, laboratory experiments involve claims and inferences about causal processes for which we have better background knowledge (of potential confounds) than in the case of 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, which is in turn likely to correlate with the maturity of the research program to which the causal hypothesis pertains or with various other general aspects that normally determine what we consider to be well-established (as opposed to tentative) knowledge. This, in turn, has little relation to whether experiments are performed in the laboratory or in the field[6].

[6] A caveat is in order, though. Some experimentalists and methodologists interpret Vernon Smith's notion of parallelism as a limited version of the notion of experimental validity: generalization is thus understood in the restricted sense of making local inferences from the experimental setting to a very limited set of situations (for a discussion of this idea, see Guala 1999). In those cases, there is a sense in which some field experiments may have achieved parallelism (or this limited sense of external validity) by the mere fact of being performed under the same situation constituting the target system, as in Guala's "mediating" experiments (2005). The majority of field experiments (e.g., see our example below), though, deal, at least in principle, with causal hypotheses and processes that go beyond the particular experimental setting and thus the external validity of their inferences applies to situations beyond the experimental setting.

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). We present this as a case study by which to illustrate that many of the ideas formulated around the notions of internal and external validity are neither accurate nor useful in the assessment of actual experiments.

## 3. 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 centres 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 centres. A flat rate fine was imposed on those parents that arrived ten or more minutes late. The other four centres, where everything was left unchanged, served as a control group. The treatment (a monetary sanction) was assigned to six of the day-care centres that were identical, in every relevant respect,[7] to the other day-care centres 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

[7] 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). Since the authors make no reference as to how this equivalence is attained, we ought to guess that it is a case of matching rather than random assignment of equivalent centers to either the treatment or the control groups.

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 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 in the field, one could say that the question on whether the purported causal relationship also holds *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.[8] The question this article poses is though: is the standard view of internal and external validity (and of the usual properties of field experiments) useful, enlightening, or true in some sense of this experiment in particular and of field experiments in general?

Upon examination, we can say that Gneezy and Rustichini's study could potentially face at least a clear threat to validity that may seem to have been overlooked

[8] 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.

by the authors. The fact that the treatment is conducted in 10 day-care centres in the same city would open the possibility of parents' learning that they are being part of an experiment. In this case, G&R´s results, in which an increase in the number of parents arriving late once a fine is introduced is observed, 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 to the parents in the control group, 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 hypothesis of 'resentful' parents. Our purpose in pointing this out is, instead, to underline the fact that the design presented and defended by G&R can be seen as potentially subject to 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 assigned to different groups impossible in practice and therefore the possibility exists for an unwanted reaction to the treatment if subjects in the treatment group feel they are being discriminated against as compared to members of the control group. 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, thus minimizing the possibility of parents learning that they are being part of a larger trial. 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 this is a field experiment. This characteristic, as Gneezy and

Rustichini underlines, 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 from what 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 reality. 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 Table 2 below), where the comparative statics of the model provide an account of the effect of changes in one variable with respect to another.[9]

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: $[(Y_2\text{-}Y_1) \neq (Y_4\text{-}Y_3)]$. The findings in each case being that $(Y_2\text{-}Y_1) > 0$ and $(Y_4\text{-}Y_3) < 0$).

[9] 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).

**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 | $X_{fine}$ | $Y_1$ | Constant |
| Control | 0 | $Y_2$ | Constant |

**G&R field experiment**

| Incomplete contract | Treatment (Putative cause: fine) | Putative effect (lateness) | Other factors |
|---|---|---|---|
| Treatment | $X_{fine}$ | $Y_3$ | Constant |
| Control | 0 | $Y_4$ | Constant |

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 in the literature could lead us to assume that G&R's experiment should not in principle be problematic: the experimenters should have no problem in justifying the proposed conclusions outside experimental conditions because the experimental and outside conditions coincide. Recall, though, that the self-declared theoretical interest behind G&R's study is not late arriving parents at kindergarten centres but a more general behavioural hypothesis regarding punishment or sanctions and deterrence. This is actually why, regarding the issue of external validity of the experiment the 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 (and 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), we cannot be certain 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 be found in southern 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. 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 experiment in 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 arriving late to one's office or other sanctionable behaviour in the workplace, the assistance rates to organizational meetings in voluntary associations, etc.). In turn, and in order to overcome the limitations of the experiment in terms of internal validity, the most obvious prescription seems to be surprisingly 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 in order to ensure that the introduction of sanctions can, in other contexts, lead to the *completion* of an underspecified or implicit contract and can ultimately act as a pricing device.

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 of external validity or vice versa. Instead, the same prescription, namely, the replication of the experiment under slight variations, is what is recommended in order to secure both its internal and external validity. In this way, a closer inspection to a concrete case study shows that many widespread ideas on external and internal validity, while resonating plausible when stated at a general level, do not fare well when applied to particular experimental instances.

## 4. 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 variations in experimental designs. In experimental economics and social science theory-driven experiments, 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 turn, and 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. 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.

This article has identified a latent debate in experimental social sciences literature regarding the relationship between internal and external validity of experiments. We adopt a critical stance to the standard position on this debate by showing that problems of either external or internal validity do not necessarily nor crucially depend neither on the artificiality of experimental settings nor on the laboratory-field distinction between experiments. More commonly, threats to internal or external validity depend on the particularities of the design and on problems with the operationalization of crucial variables, and there seems to be no grounds to posit a general trade-off between the internal and external validity of experiments.

In view of this, what perhaps needs an explanation instead is the fact that references to the trade-off are so abundant in the methodological and philosophical literature around experiments, and in particular, in experimental economics. We hypothesize that what is very often at play in the literature is the conflation of two different ideas: that extrapolating from experimental results is seldom easy or straightforward and that there is a trade-off between internal and external validity. Clearly, there is very often the possibility that once a causal link between two variables has been established experimentally, we might not know (depending on our background knowledge of the process under study) the conditions under which that causal link may manifest in the outside world because many other intervening factors may be present under "real life" conditions. This surely happens time and again, both in the natural and in the social sciences, but to equate this phenomenon to a trade-off between internal and external validity seems rather odd: in most of the instances, it is because we have confidently and experimentally established a causal relationship between two variables that we feel then enabled to try and extrapolate the findings to other settings, where other intervening variables may also operate.

Put in this way, the shielding of experimental conditions from extraneous factors is to be seen as a first step in achieving reliable knowledge that can allow for latter generalizations about causal relations across other, varied, circumstances. It would seem too demanding of any scientific practice to expect that comprehensive knowledge about a causal relation of interest would stem from a single exercise: in this case, from one particular experiment. Instead, scientists aim (most of the times collectively) at gradually accumulating robust knowledge about causal relations by performing an array of different yet related experiments and by combining them with theoretical reflection, observational data, and any other type of available evidence. Because we normally do not aspire at finding out (with just one experiment) about all the circumstances in which a potentially causal link holds or all the details on the causal mechanism concerned, we typically, thus, do not consider that an experimenter devoted to introducing small variations in the experimental environment of a well-established causal link is performing trivial work.

Restating the series of platitudes about scientific practice mentioned just above has only one purpose: we feel compelled to insist on the fact that these well-known facts be best not confused with or referred to as the *trade-off between internal and external validity of experiments*. If a trade-off between the internal and external validity of

experimental inferences operates it must mean something other than a mere reference to the difficulties of obtaining exhaustive and definitive knowledge about causality. We have tried to trace these other possible meanings in the current methodological literature. To the extent that we have failed to find compelling accounts of those other meanings of the trade-off between internal and external validity of experiments we propose the abandonment or reformulation of this (as in the words of Cartwright, op.cit) "well-known methodological truism".

**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.

Cartwright, N. 2007. Hunting Causes and Using them. Cambridge: Cambridge University Press.

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. 1999. The problem of external validity (or "parallelism") in experimental economics. Social Science Information 38; 555.

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

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

Guala, F. 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

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.

Levitt S.D. and List, J. 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* 21:153-74.

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

Lusk J.L., Pruitt J.R., Norwood B. 2006. External validity of a framed field experiment. *Economics Letters* 93: 285-290.

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*. Cambridge: Cambridge University Press.

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.

Shadish, W.R., Cook, T.D., & Campbell, D.T. 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

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. Heijine, 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.