03228

(Quasi- and Field) experiment, history of

Alex Díaz

María Jiménez-Buedo

David Teira (**Corresponding author**)

Alex Díaz

Dpto. de Lógica, Historia y Filosofía de la ciencia. UNED

Paseo de Senda del rey 7 | 28040 madrid

alex.diaz@bec.uned.es | (34) 91 398 83 92

María Jiménez-Buedo

Dpto. de Lógica, Historia y Filosofía de la ciencia. UNED

Paseo de Senda del rey 7 | 28040 madrid

mjbuedo@fsof.uned.es | (34) 91 398 83 92

David Teira

Dpto. de Lógica, Historia y Filosofía de la ciencia. UNED

Paseo de Senda del rey 7 | 28040 madrid

dteira@fsof.uned.es | (34) 91 398 83 92

Abstract

Field trials and quasi-experiments are comparative tests in which we assess the effects of one intervention (or a set thereof) on a group of subjects as compared to another intervention on another group of similar characteristics. The main difference between field trials and quasi-experiments is in the way the interventions are assigned to the groups: in the former the allocation is randomized whereas in the latter is not. We are going to see first the different roles played by randomization in medical experiments. Then we discuss how controlled field trials, originating in psychology, spread to the social sciences throughout the 20th century. Finally, we will show how the idea of a quasi-experiment appeared around a debate on what constitutes a valid test and what sort of controls guarantee it.

Keywords

Field trial, quasi-experiment, randomization, validity, causality, controls, bias, Donald Campbell

1. Field experiments and Quasi-experiments

Field experiments and quasi-experiments are comparative tests in which we assess the effects of an intervention (or a set thereof) on a group of subjects as compared to another intervention on another group of similar characteristics. The main difference between field trials and quasi-experiments is in the way the interventions are assigned to the groups: in the former the allocation is randomized whereas in the latter is not. For more than a century now, there is an ongoing debate about what sort of methodological controls in the comparison warrant causal inferences in such experiments. In particular, it is under discussion in what sense randomized allocations provide superior grounds for causal inferences. We are going to see, in the first place, the roles played by randomization in medical experiments. Then we will discuss how controlled field experiments, originating in psychology, spread in the social sciences throughout the 20th century. Finally, we will show how the idea of a quasi-experiment appeared around a debate on what constitutes a valid test and what sort of controls guarantee it.

2. Randomized experiments

Randomized experiments originate mainly in medicine, where the fair comparison of therapies is well documented centuries ago. As early as in 1662, the Flemish physician J.B. Van Helmont suggested settling a dispute about the efficacy of bloodletting and purging with a randomized allocation of treatments: "Let us take out of the hospitals […] 200 or 500 poor people, that have fevers, pleurisies. Let us divide them into halves, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation; but you do, as ye know […].We shall see how many funerals both of us shall have". A 1747 test of six remedies for scurvy is considered the first actually controlled trial: conducted by James Lind, a naval surgeon, it showed the superior efficacy of (vitamin-C-containing) oranges and lemons as compared to the other alternatives (against the opinion of both the Royal College of Physicians and the Admiralty). As Iain Chalmers (2006) has extensively argued, randomization originated in comparative experiments as a method for the correction of biases in the allocation of the tested treatments. If randomized, the preferences of the experimenter allocating them would not contaminate the comparison (e.g., assigning her favorite treatment to one particular group of patients). Randomization was just one possible *control* for the selection bias; an alternative one was alternating treatments between patients. Medical experimenters identified many other biases in their tests, developing *controls* for them (e.g., blinding as early as 1784). As it often happens with error-correction procedures in most scientific fields, these debiasing methods were grounded in the experimenters' experience, without any substantial theory to account for their efficacy. In the first decades of the 20th century, medical trials implementing all sorts of controls were common, but the standardization of the randomized clinical trial (RCT) began only with the 1947 test of Streptomycin, an antibiotic drug, commissioned to Austin Bradford Hill, a medical statistician, by the British Medical Research Council (Teira 2013a).

Hill relied on the foundations for experimental design developed by the great statistician Ronald Fisher for agricultural research on fertilizers and soils. Fisher used randomization (now explicitly defined on the probability of receiving a given treatment) to ground significance testing: how likely it is to observe an experimental outcome equally or more extreme than the one actually obtained in an infinite series of repetitions of the trial? Randomization secured that the actual allocation of treatments

was just one random draw in that series so we could calculate the *p*-value for the actual outcome obtained on the basis of the distribution of all outcomes for every possible allocation. Furthermore, Fisher defended the role of randomization in causal inference: the interference of potential confounding factors could be disconnected from the treatment outcome in a series of replications of the experiment if the allocation was randomized.

Bradford Hill merged these three roles of randomization (debiasing, significance testing, causal inference) in the trial of Streptomycin, establishing the first template for the RCT. The pharmaceutical revolution of the 1950s also brought about the expansion of drug testing, and the template was gradually refined with further debiasing controls (e.g., double blinding) and statistical tools (e.g., power calculations, showing the adequate sample size for a statistically sound test). In 1962, RCTs for safety and efficacy became compulsory for the regulatory approval of new drugs by the American Food and Drug Administration. This created a booming industry of pharmaceutical trials still alive today, although under serious discussion (Goldacre 2012).

3. Field experiments in the social sciences

The history of field experiments in the social sciences is still to be written. The few existing accounts have not so far been inclusive of all the social scientific disciplines, and this leaves us with a fragmented portrayal of experimentation in the social science. In the case of Economics, for instance, and according to Steven Levitt and John List (2009), the use of field experiments would originate in the early 20th century, growing gradually to an upsurge in the last decade (the 2000s). Here we will motivate the importance of two questions that arise in the light of this yet fragmentary history: how did social scientists reach a consensus on the experimental design of field trials? Why did they use this design so scarcely until recently? The methodological design and interpretation of a field trial is in itself far from obvious: in principle, we are trying to grasp the causal effect of an intervention comparing it with an alternative one on two groups of subjects, both entirely alike except for the effect of the intervention. We find explicit methodological discussions on such comparisons in the 19th century (e.g., Mill's method of difference), but experimenters need to agree, on the one hand, on the controls they impose on the two groups in order to make sure that no confounding factors make a difference between them. And, on the other hand, if there is a quantitative difference between the outcomes, experimenters need to agree also on a statistical procedure to assess the effects.

Historians of statistics, following Theodore Porter (1995), have shown how the use of statistical techniques in public policy was presented as an alternative to expert judgment. RCTs in medicine came to replace the opinion of individual doctors about the efficacy of treatments (sometimes based on laboratory evidence, often just on their own clinical experience). Of course, many physicians considered their judgment objective enough to ground the assessment. And, furthermore, a comparative approach was often considered unethical since it would deprive a group of patients of a treatment that a physician would have recommended otherwise. This is why the agreement on a comparison constrained by a series of controls (often on the physician administering the treatment) was only reached after several decades in the 1970s, when the first meta-analyses started ranking trials for the aggregation of evidence (for an instance of such lists of controls, see Bero & Rennie 1996).

As to the assessment of the comparison, right from the beginning there were serious divergences on what were the proper statistical grounds to judge the difference between

interventions (Teira 2011). The controversy between Fisher and Jerzy Neyman in the 1930s is well documented, for instance: both parties shared a frequentist conception of probability, but disagreed on how to apply it to an experimental test of a given hypothesis. Fisher wanted to rely exclusively on p-values: assuming the truth of the hypothesis that there is no difference between treatments, if we observe a substantive divergence between the two groups, either an exceptionally rare chance has occurred or the hypothesis is not true. But the data alone cannot establish whether the former or the latter is the case (or whether both are). Neyman, together with Egon Pearson, developed a different rationale for the testing of hypotheses: instead of assessing the plausibility of a single (null) hypothesis, we should have a criterion for choosing between alternative exclusive hypotheses, with a known probability of making the wrong choice in the long run. Fisher considered this latter approach suitable only for industrial quality controls, but epistemically ungrounded. Neyman opened the way instead to a decision-theoretic appraisal of hypothesis testing. Eventually it led to a full-fledged Bayesian alternative, already emerging in the 1960s and thriving today, in which RCTs can be designed and interpreted without neither randomization nor p-values.

Hence, the history of field trials cannot be written, as Leavitt and List suggest, as if once Fisher statistically grounded the design (randomization plus significance testing), it just spread to the medical and socials sciences without controversy. For instance, Stephen Ziliak (2014) shows that even randomization was controversial in proto-economic trials: already in the 1910s the pioneer of statistics, *Student* (the pen name of William Gosset) defended a balanced design of agricultural plots as an alternative to randomization, since the confounding factors are not independent and identically distributed.

Historically, field trials emerged among many other approaches to social observation (see, e.g., Maas & Morgan 2012 regarding economics) and the different controls in their experimental design were tried one by one in a variety of settings. According to Stephen Stigler (1992) and Ian Hacking (1988), the philosopher Charles S. Peirce, together with his student JosephJastrow, conducted the first sound randomized experiment in psychology between 1883 and 1884. It tested Fechner's hypothesis about the existence of a perception threshold below which one cannot discern small differences. The experimenter presented different weights to the subject, ordering them according to the cards drawn from a standard deck. Peirce used this randomization scheme as a debiasing procedure (in order to control the subject's guesses about weights), without any explicit statistical implication. Yet the procedure was not adopted in subsequent experiments on the topic, although it was actively used in all sorts of experiments on telepathy for several decades.

According to Trudy Dehue (1997), by the end of the 19th century out of Fechner's psychophysics research program emerged a causal approach to the study of learning transfer (subjects were asked to perform an ability task with either the left or the right hand to measure gradual performance improvement and then they were tested to see whether the trained skill had transferred to their other hand). This causal approach, of which randomization of treatments was already an integral part, was gradually adopted by psychologists of education, led by the Columbia psychologist Edward L. Thorndike, and eventually came to be increasingly implemented at a major scale in American educational research. Here, still according to Dehue, the Stanford psychologist John Edgard Coover presented the control group in 1907 as a methodological device to correct a confounder in training experiments: anyone who takes the same test twice is likely to do better the second time –on top of the multiple factors that may affect the subject in between. In the 1910s, control groups were used in experimental educational

research conducted in schools on the most effective means to teach. In the 1920s we find a variety of comparative studies with different controls (Oakley 2000, pp. 164-168) and different degrees of quantitative analysis. For instance, after graduating with Thorndike in Columbia, William McCall (1923) started testing the validity of classroom practices with control groups, using randomization as an allocation procedure.

The 1920s were the age of quantitative research in American social science (Porter & Ross 2003, Oakley 2000, pp. 163-197). Although educational research was gradually left behind, psychology became a science of controlled experiments, with controls intending to warrant bias-free causal analysis, favoring the laboratory to the field. This is how it features in the methodological handbook of the Social Sciences Research Council (SSRC 1932)a major funding body of American social science at that point. Statistics, and in particular statistical economics, dealt instead with experimentally uncontrolled factors. Nonetheless, other social scientists tried to do controlled field research: still in Chicago and under the influence of Charles Merriam, the political scientist Harold Gosnell (1927) tried to find a third way studying the factors influencing voter turnout on 6000 Chicago citizens. Gosnell divided them into an experimental and a control group organized by place of residence; the former received notices urging them to register and vote, and their turnout was followed up in elections in 1924 and 1925.

The statistical techniques implemented in these early field experiments (often Pearson's correlation theory) did not contribute much to differentiate between true and spurious causes (Turner 2007). The weight of causal inference still lied on the controls, where randomization was not always the most prominent. E.g., another Columbia graduate, the sociologist Francis Stuart Chapin (1947), introduced the term *ex post facto experiment* to name studies where the conditions are not manipulated but selected by the investigator and the comparison is controlled by matching groups –considered a licit alternative to randomization.

Fisher's statistical foundations were explicitly incorporated into the design of these trials already in the 1940s (e.g., Lindquist 1940). Yet, in fields such as educational research this approach apparently did not yield significant results stimulating further research (Oakley 2000). However, field trials in the social sciences did not take off until the 1960s (Dehue 2001). Just as it happened with RCTs in medicine, its success owes much to American regulatory pressure: several Acts enacted by Congress in the early 1960s mandated evaluation of the programs implemented, just as the 1962 Food and Drug Administration required "adequate and well-controlled clinical studies" for proof of efficacy and safety of new treatments. However, whereas in the three following decades thousands of RCTs would be conducted, in the social sciences there were only a few hundreds. A paradigmatic field trial of this "golden age of evaluation" (Oakley 2000) is the New Jersey Income Maintenance Experiment (NJIME), conducted from 1968 to 1972, with nearly 1500 participant households at a cost of $7.8 million (1971). Based on the doctoral dissertation of Heather Ross (1970), the trial tested a welfare policy suggested by the economist Milton Friedman: a negative income tax, by which people earning below a certain amount would receive supplemental pay from the government instead of paying taxes to the government. The general intuition is that, unlike indirect assistance schemes, this negative tax would give people incentives to work and sustain themselves. The NIJME tested whether different combinations of income guarantees plus a tax rate on other income had an effect on the work effort of the participants or prompted any other lifestyle changes. This was all followed by questionnaires every three months. The initial assessment of the outcome concluded that

work effort did not decline in the treatment group (measured by the number of hours of employment), although further analyses of the data pointed out that there was indeed a reduction. The negative income tax worked better in this respect than some welfare schemes though, but it did not outperform others. The NIJME had a control group and treatment allocation was randomized, but a number of potential confounders and biases were soon detected (Greenberg & Shroder 2004, 194-195): e.g., misreporting of earnings or misrepresentation of the sources of income of families. Many other trials dealt with a variety of policies and social settings, with different degrees of success –for a review see Oakley 2000, pp. 198-230; Greenberg & Shroder 2004.

During this golden age of trials in the 60s and 70s, a new discipline, *evaluation research*, gradually emerged. In close collaboration with administration researchers, truly interdisciplinary groups of social scientists coming from different disciplines, like economists, sociologists, educational researchers and psychologists, participated in evaluating the different aspects of implemented policies and programs, making widespread use of both experimental and non-experimental methods (Alkin 2012). Central to this team of new policy evaluators was Donald T. Campbell, a Berkeley graduate in psychology, with broad interests in the social sciences, philosophy, and the methodology of causal inference. The publication of Campbell and Stanley's book in 1963 (republished in 1966 under a more general heading) was a major breakthrough in methodological research for the social sciences (experimental and else), and it still is among the most widely cited social scientific methodological works. As we will see, Campbell and his collaborators helped to create a unified language for policy evaluators in their analysis of both experimental and non-experimental data and their work was instrumental in the listing and systematization of the possible biases that can threaten experiments, some of which are specific to experimentation in the social sciences.

From a statistical standpoint, field trials reached maturity, integrating full-fledge hypothesis testing and power calculations –although the controversy on its proper interpretation has not completely vanished: it is still open to discussion under which circumstances statistical significance provides good enough grounds to interpret the test of hypotheses (e.g., Ziliak & McCloskey 2008).

The experience accumulated in this golden age showed a number of biases that might interfere in the outcome of the experiment. Levitt and List (2009) list, for instance, the following: *randomization bias*, by which the allocation process would exclude risk averse participants; *attrition bias*, participants drop out of the experiment more in one of the groups, biasing the comparison; the *Hawthorne effect*, participants change their behavior because they know they are being observed in an experiment; *substitution bias*, by which the participants in the control group violate the protocol seeking alternatives to the treatment they receive. The problem with these biases is that there might not be an obviously effective debiasing procedure (e.g., Teira 2013b). In other words, the list of controls that would bring about a consensus on the outcome of trial is still under discussion.

In recent years, a third generation of field trialists has emerged with great strength in economics. Led by the Abdul Latif Jameel Poverty Action Lab (J-PAL) at the MIT, a cohort of development economists are testing social interventions in third world countries, where the costs of implementing such programs are significantly lower. The randomization of treatment often takes place in naturally occurring settings, and the participants often remain unaware of the experiment conducted (correcting thus the randomization bias and the Hawthorne effect). Trials are often designed for theoretical purposes (theory testing or parameter measurement), but their current popularity owes

much to their promise of delivering evidence-based policies to fight poverty. In this respect, trialists such as Esther Duflo (2010) advocate the necessity of an evidence hierarchy for policy making: such as physicians use evidence based rankings to decide on prescriptions according to the best available data (with meta-analyses and RCTs on top of the pyramid), social scientists should ground their recommendations according to the same principles. Are field trials as reliable regarding policies as RCTs with drugs? The future of the field will probably depend on the answer social trialists deliver in the forthcoming years (Teira & Reiss 2013).

4. Quasi-experiments

Although an enthusiast and public advocate of social experimentation, Campbell was well aware of the fact that randomization was often not feasible in the social realm (due to ethical, administrative, or pragmatic obstacles). With this in mind, Campbell developed the idea of quasi-experiments and developed a system of different designs that fell under the category (Campbell & Stanley 1963). Although the term quasi-experiment had previously been used by other authors, Campbell fixed and systematized its use to the current one. Quasi-experiments share with all other experiments a similar purpose (to test descriptive causal hypotheses about manipulable causes) as well as many structural details, such as control groups and pretest measures. But, by definition, quasi-experiments lack random assignment of groups to treatments. Assignment to the different interventions can occur either by means of participants self-selection, or alternatively, by means of administrator selection, by which teachers, bureaucrats, or researchers, decide who gets what treatment. In the latter case, quasi-experimenters may still have a considerable control over how the assignment is executed.

What gives quasi-experiments their close-to-experimental nature is that the cause is manipulable and occurs before the effect is measured. However, since the allocation to conditions is not performed at random, the inferences that can be made from results are typically limited, because we cannot rule out the existence of systematic differences between groups originating in a confounding factor. These differences will often constitute alternative explanations for the observed effect, in direct competition with the tested treatment. Researchers, though, may explore these alternative explanations and rule some or even all of them through other alternative tests either *ex ante* or *ex post*. However, in cases in which this potential list of alternative explanations is very long, quasi-experimental results may not be conclusive. Quasi-experimental designs need to be very carefully planned in order to minimize the potential threats that stem from the lack of randomization. For this reason, quasi-experiments tend to be more complex than true experiments both in their design and in their analysis (Shadish et al. 2001).

One of the most important outputs of Campbell's methodology is his approach to the question of experimental validity, an approach he transposed from previous concepts of measurement validity in psychology into social scientific experimentation (Heukelom 2011). Campbell first introduced the terms in his 1957 article "Factors Relevant to the Validity of Experiments in Social Settings". In this first definitional attempt internal validity was seen as the basic minimum requirement of an experiment, expressed by reference to the following question: "did in fact the experimental stimulus make some significant difference in this specific instance?"; and external validity, in turn, referred to the question of representativeness, or generalizability: "to what populations, settings, and variables can this effect be generalized?"(1957, p.297). Though certain aspects of Campbell's validity typology have been controversial since the beginning, the distinction between internal and external validity remains a central tenet of the

contemporary conceptual assessment of experimental validity (Jiménez-Buedo & Miller, 2010).

After this initial definition, Campbell, together with a number of collaborators gradually reformulated the distinction between internal and external validity and gradually added more validity types into what is now a stable four-fold validity typology: *statistical conclusion validity*, *internal validity*, *construct validity* and *external validity*. Statistical conclusion validity has been defined as the validity of inferences about the correlation between treatment and outcome. Construct validity refers to the inferences about the higher order constructs that represent sampling particulars (Shadish et al. 2001). These latter two concepts, though common currency in social psychology and related fields in the social sciences, are much less used in more recent experimental fields, like behavioral economics. Often, the first two concepts (statistical conclusion and internal validity) are usually grouped together, whereas construct and external validity are also often conflated, suggesting, in accordance to Campbell himself (Shadish et al. 2001), that all of the categories can be subsumed under a broader, more encompassing, internal/external pair.

Campbell developed this methodological discussion as a reaction against an unreflective use of randomization in social psychology, as if it were the only necessary control to secure the validity of an experiment. Through the years, the validity typology helped to identify and analyze a series of common sources of biases, or threats to validity. Examples of threats to internal validity are phenomena like *selection* (of units to treatments), *history* (i.e., the existence of concomitant events to the treatment that may be actually responsible for the observed effect), *differential attrition* (or mortality) of subjects in the different experimental groups, or *maturation*, where naturally occurring changes over time could be confused with a treatment effect. All of these are confounding factors that can be mistaken for effects of the implemented treatment. In turn, threats to external validity usually originate in confounders that may lead us to infer erroneously that a causal relation will hold over contexts other than the one under study. Examples of such threats are the interaction of selection and treatment (where the relation between *x* and *y* will not hold beyond the type of subject who took part in the experiment), or interaction of setting and treatment (where the relationship between *x* and *y* will only hold in the experimental setting). Knowledge about these "threats" was, for Campbell and his collaborators, a matter of practice: detecting ways in which experimentalists were arriving at wrong inferences from looking at the results of their experiments was thus for Campbell a collective, inductive process and very importantly, relative to the details of a particular study. Therefore, the lists of validity threats never intended to be exhaustive, and many additions (and substractions, and reclassifications) were made over the years.

The list of threats is still to date routinely used for those purposes by field social experimentalists that contrast it, when choosing experimental designs, with feasibility or budgetary constraints. Campbell though was always careful to remind us that performing these checks did not safeguard any design from a potentially infinite list of threats. In this sense, Campbell showed some concern and went as far as considering the lists potentially "dangerous if they led readers" to overestimate their capacity as shields from error (1986, p. 154). This overestimation, for its practical implications, seemed a bigger concern to Campbell than the fact, of which he was also aware, that his distinction between internal and external validity had been often misunderstood in ways that portrayed it as having to do with the distinction between field experiments and pure laboratory research (Jiménez-Buedo, 2011).

As regards the choice between lab versus field experiments, there has been, historically, a vivid debate concerning the advantages of both in social psychology. Among the early supporters of field experiments we find Kurt Lewin (e.g., 1939), who highlighted the need to focus on the balance of forces influencing behavior in social situations, developing methods to perform experiments in everyday life scenarios. During the 20th century, psychology mostly focused on controlled laboratory experiments, but in the 1960s, authors like McGuire (1969) vindicated again field experiments in social psychology from a methodological standpoint.

The enthusiasm for the field and immersive experimental settings in the 1960s also yielded Stanley Milgram's and Philip Zimbardo's famous experiments on obedience and interpersonal dynamics. Milgram (1963) reproduced a clinical scenario to observe obedience to authority: Participants were instructed to administer electrical shocks to other participants and witnessed the (pretended) consequences of pain and disgust. Zimbardo and cols. reproduced the structure of a prison in the basements of a Stanford building, and instructed participants to behave like wards and prisoners, which resulted in a hostile atmosphere as participants followed the instructed roles (see e.g., Haney et al., 1973). Such extreme social interactions somehow dissuaded other psychologists from following this experimental path and, as of today, social psychology still hinges more on the laboratory than on the field (Savoley & Williams-Piehota 2004).

5. Cross-references

Experimenter and Subject Artifacts: Methodology
External validity
Hypothesis Testing, Methodology and Limitations.
Internal Validity
Nonequivalent group designs
Random Assignment: Implementation in Complex Field Settings
Selection bias, statistics of
New Approaches to Reliability and Validity Assessment
Experimental Design, Overview
Optimal Experimental Design
Experimental Design: Randomization and Social Ex.
Quasi-Experimental Designs
Field Experiments
Natural Experiments
Experimental Economics
Experimentation, history of (in psychology)

6. References

ALKIN, M. C. 2012. *Evaluation roots : a wider perspective of theorists' views and influences.* Los Angeles, CA: SAGE Publications.

BERO, L. A. & RENNIE, D. 1996. Influences on the quality of published drug studies. *International Journal of Technology Assessment in Health Care,* 12, 209-237.

CAMPBELL, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin,* 54, 297-312.

CAMPBELL, D. T. & STANLEY, J. C. 1963. Experimental and quasi-experimental designs for research on teaching. *In:* GAGE, N. L. (ed.) *Handbook of Research on Teaching.* Boston, MA: Houghton Mifflin.

CAMPBELL, D. T. & STANLEY, J. C. 1966. *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.
CHALMERS, I. 2006. Why fair tests are needed: a brief history. *Evidence Based Medicine,* 11, 67-68.
CHAPIN, F. S. 1947. *Experimental designs in sociological research.* New York, NY: Harper.
DEHUE, T. 1997. Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis,* 88, 653-673.
DEHUE, T. 2001. Establishing the experimenting society: the historical origin of social experimentation according to the randomized controlled design. *American Journal of Psychology*, 114, 283-302.
DUFLO, E. 2010. *La politique de l'autonomie.* Paris: Seuil.
GOLDACRE, B. 2012. *Bad Pharma*, London: Fourth State.
GOSNELL, H. F. 1927. *Getting out the vote; an experiment in the stimulation of voting*, Chicago, IL: The University of Chicago Press.
GREENBERG, D. H. & SHRODER, M. 2004. *The digest of social experiments*. Washington, DC: Urban Institute Press.
HACKING, I. 1988. Telepathy: Origins of randomization in experimental design. *Isis*, 79, 427-451.
HANEY, C., BANKS, W. C., & ZIMBARDO, P. G. 1973. Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, 1, 69–97.
HEUKELOM, F. 2011. How validity travelled to economic experimenting. *Journal of Economic Methodology*, 18, 13-28.
JIMÉNEZ-BUEDO, M. 2011. Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*, 18, 271-282.
JIMÉNEZ-BUEDO, M. & MILLER, L. M. 2010. Why a trade-Off? The relationship between the external and internal validity of experiments. *Theoria*, 25, 301-321.
LEVITT, S. D. & LIST, J. A. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53, 1-18.
LEWIN, K. 1939. Field theory and experiment in social psychology: concepts and methods. *American Journal of Sociology*, 44, 868-896.
LINDQUIST, E. F. 1940. *Statistical analysis in educational research,* Boston, MA: Houghton Mifflin.
MAAS, H. & MORGAN, M. S. 2012. The observation and observing in economics. *History of political economy*, 44, 1-24.
MCCALL, W. A. 1923. *How to experiment in education,* New York, NY: The Macmillan company.
MCGUIRE, W. J. 1969. Theory-oriented research in natural settings: the best of both worlds for social psychology. In SHERIF, M. & SHERIF, C. W. (eds.) *Interdisciplinary relationships in the social sciences*. New Jersey, NJ: Transaction Publishers, 21-51.
MILGRAM, S. 1963. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67, 371-378.
OAKLEY, A. 2000. *Experiments in knowing: gender and method in the social sciences.* New York, NY: New Press.
PORTER, T. M. & ROSS, D. 2003. *The Cambridge history of science, vol. 7: The Modern Sciences.* Cambridge, MA: Cambridge University Press.
PORTER, T. M. 1995. *Trust in numbers: the pursuit of objectivity in science and public life.* Princeton, NJ: Princeton University Press.

ROSS, D. 1991. *The origins of American social science.* Cambridge, MA: Cambridge University Press.

ROSS, H. L. 1970. *An experimental study of the negative income tax.* Doctoral Dissertation. MIT.

SALOVEY, P. & WILLIAMS-PIEHOTA, P. 2004. Field experiments in social psychology. *American Behavioral Scientist,* 47, 488-505.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. 2001. *Experimental and quasi-experimental designs for generalized causal inference,* Boston, MA: Houghton Mifflin.

SOCIAL SCIENCE RESEARCH COUNCIL. COMMITTEE ON SCIENTIFIC METHOD IN THE SOCIAL SCIENCES, RICE, S. A., SECRIST, H., MACIVER, R. M. & LASSWELL, H. D. 1932. *Methods in social science, a case book compiled under the direction of the Committee on scientific method in the social sciences of the Social science research council.* Chicago, IL: The University of Chicago press.

STIGLER, S. M. 1992. A historical view of statistical concepts in psychology and educational research. *American Journal of Education,* 101, 60-70.

TEIRA, D. 2011. Frequentist versus Bayesian clinical trials *In:* GIFFORD, F. (ed.) *Philosophy of Medicine.* Amsterdam: Elsevier, 255-297.

TEIRA, D. 2013. Blinding and the non-interference assumption in medical and social trials. *Philosophy of the Social Sciences,* 43, 358-372.

TEIRA, D. 2013. On the impartiality of early British clinical trials. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 44, 412-418.

TEIRA, D. & REISS, J. 2013. Causality, impartiality and evidence-based policy. *In:* CHAO, H.-K., CHEN, S.-T. & MILLSTEIN, R. (eds.) *Mechanism and causality in biology and economics, vol. 3.* London: Springer, 207-224.

TURNER, S. P. 2007. Defining a discipline: Sociology and its philosophical problems, from its classics to 1945. *In:* TURNER, S. P. & RISJORD, M. W. (eds.) *Philosophy of anthropology and sociology, First ed.* Amsterdam: Elsevier, 3-69.

ZILIAK, S. 2014. Balanced versus randomized field experiments in Economics: Why W. S. Gosset aka "Student" matters. *Review of Behavioral Economics,* 1, 167-208..

ZILIAK, S. T. & MCCLOSKEY, D. N. 2008. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives.* Ann Arbor, MI: University of Michigan Press.