CERN IdeaSquare Journal of Experimental Innovation, 2025; 9(3):4-10
DOI: https://doi.org/10.23726/cij.2025.1794

METHODOLOGICAL NOTE

Laboratory experiments in innovation research

Eric Guerci[1]

[1]Université Côte d'Azur, CNRS, GREDEG, 250 rue Albert Einstein, CS 10269, 06905 Sophia Antipolis Cedex, France
Corresponding authors: eric.guerci@univ-cotedazur.fr

WHAT ARE LABORATORY EXPERIMENTS?

Laboratory (lab) experiments combine two Latin roots, *laborare* (to work) and *experiri* (to try or to test), which together capture the very essence of innovation studies. They may thus be rephrased as "working practices based on testing". The first documented use of the term "laboratory" dates back to the late seventeenth century referring to alchemist's workshop (Morris 2021), where early "innovation researchers" tested systematically physical and chemical phenomena in the pursuit of innovative discoveries. In the same century, modern experimental science was emerging through controlled experimental setups approximating laboratory conditions, as illustrated by Galileo Galilei's inclined-plane studies of gravity or by classical optical experiments, such as Christiaan Huygens' double-refraction experiment and Isaac Newton's prism experiment.

The implicit guiding principle of such causal inquiries was the *ceteris paribus principle* (Hacking, 1989): the idea that, by isolating and manipulating a specific causal factor (e.g., slope of the inclined plane) while holding other relevant conditions fixed through experimental control (same body, surface, track length, …), stable quantitative relations (e.g., distance–time relationship) could be identified and thus laws between physical variables could be uncovered (e.g., law of uniformly accelerated motion). This principle is still largely implementable in hard-science experiments (physics and chemistry), where closed systems can be investigated by restricting the predominant sources of uncertainty to measurement and control errors. For such natural phenomena, quantitative modeling rests on a relatively viable deterministic assumption, where uncertainty is confined to exogenous variables often treated as negligible measurement errors (Hacking, 1989; Pearl, 2009). Many innovation scholars, particularly in engineering and design research, conduct laboratory experiments to test prototypes, evaluate technical artifacts, or assess system performance by constructing controlled (implicitly *ceteris paribus*)[1] and iterative processes in which specific design parameters are deliberately varied while others are held constant (Thomke, 1998). *These early ambitions and experimental approaches, originating in the earliest laboratory-like practices, remains valid and continues to support a part of innovation research*, as it did in paradigmatic innovation laboratories such as the Menlo Park and West Orange laboratories of Thomas Edison (Pretzer, 1989) and extends to more recent participatory innovation environments such as fab labs and makerspaces (Browder et al., 2019; Gershenfeld, 2005).

It is worth highlighting so far two essential dimensions that will accompany our discussion.

1. Since the early instances described here above, *lab experiments are not purely observational activities* aiming at detecting associations between facts or variables, but *empirical tools designed to identify causal effects or relationships between them*. Indeed, innovation studies is inherently concerned with causal processes, as innovation refers to the generation of novel outcomes through specific mechanisms, interactions, and institutional conditions (Boudreau & Lakhani, 2016).

2. (To advance a perspective that will be central to the remainder of this paper) early laboratory-like practices, such as the above-mentioned alchemical ambitions, were largely use-oriented toward *evaluative discovery*, that is, identifying whether specific causes could entail reproducible effects (such as particular material transformations) and thus be exploited. In contrast, early scientific laboratory experiments were driven by an *epistemic* ambition to understand the causal structure underlying natural phenomena. This dichotomy persists in innovation research through the distinction between policy-oriented vs mechanism-oriented (Boudreau & Lakhani, 2016). The evaluative research paradigm focuses on assessing outcomes or effectiveness (e.g., policy impacts), whereas epistemic research aims at generating new knowledge by developing concepts, metrics, or by identifying underlying mechanisms (processes) that enable new understanding (e.g., novel indicators or societal mechanisms).

[1] It is worth noting the growing trend in engineering research towards the use of fully virtual test-bed solutions, namely *digital twins*, to simulate *ceteris paribus* conditions in controlled virtual environments prior to physical prototyping (Fuller et al (2020)).

In this paper, we focus on the scientific literature showing how lab experiments can address the need to *understand causally social and human phenomena*, by combining the systematic observation of human behavior with the controlled manipulation of the environment. In line with this perspective, it is worth recalling the major methodological breakthrough in experimental science brought about by the work of Ronald A. Fisher widely regarded as father of modern statistics and experimental design. Indeed, his work introduced *statistical methods as essential tools to quantify uncertainty and support causal inference* in contexts characterized by heterogeneity among experimental units and uncontrollable variability in the systems under study, such as biology and medicine. He introduced what is still regarded today as the gold standard of experimental design in domains characterized by irreducible variability, such as biology, medicine, and the social sciences: the randomized controlled trial (RCT). An RCT is an experimental design that randomly assigns units to treatment and control groups, thereby enabling causal inference by ideally removing the influence of confounders. More generally, Fisher's revolutionary contribution lies in his proposal to *use statistical methods to quantify uncertainty and support inferential conclusions by explicitly accounting for variability, while achieving control over sources of uncertainty primarily through experimental design*.

We introduce thus a modern and commonly adopted definition of a lab experiment in innovation research applied to the study of social systems: a set of monitored activities such as tasks and questionaries designed and conducted according to a strict set of procedures and materials including recruitment strategies, instructions and other operational elements (*the protocol*), with the aim of establishing that the observed effects (*the outcomes of the main activity*) are caused by the manipulated independent variables (*treatments*), rather than by alternative factors (*potential confounders*). Treatments are obtained as variations of the protocol; they can therefore be characterized by simple modifications, among others, in the instructions, the recruitment strategy or the materials. Potential confounders are variables (either observed or unobserved) that are correlated with both the treatment and the outcome, i.e., they generate a further uncontrolled variation in the outcome that is not attributable to the treatment. In lab experiments, randomization is the gold standard because assigning participants randomly to treatment conditions ensures that potential confounders are, on average, equally distributed across groups, thus breaking any systematic association between confounders and treatment. The laboratory ensures a controlled environment in which the protocol can be implemented and reproduced consistently.

Figure 1 presents, for illustrative purposes, a configuration of a minimal laboratory experiment based on a protocol with a single task, a binary treatment and a binary outcome, used to test whether consuming hot beverages or warm foods influences the preference for a mountain or beach holiday. A plausible confounder could be an unintentionally recruiting of students immediately after intense physical activity. This factor may affect (correlate with) both the holiday preference (beach-relax vs mountain-effort associations) and the enjoyment of drinking hot beverages or eating warm foods. Random assignment breaks the association between the confounder and the treatment, thereby preventing confounding bias.

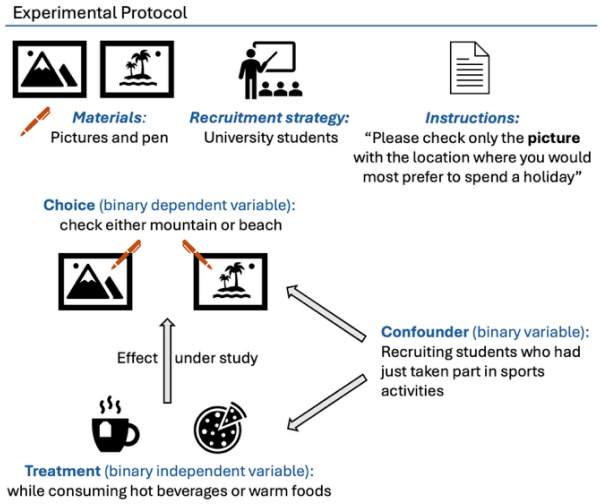

Fig.1. Illustrative example of a laboratory experiment.

WHAT ARE LAB EXPERIMENTS GOOD FOR?

The ability to make a causal claim within an experiment has a specific name and is usually referred to as *internal validity*. This notion refers exclusively to the degree to which an experiment allows us to confidently conclude that the observed effect is caused by the manipulated independent variable, rather than by alternative factors. High internal validity is supported by replicable and tightly controlled experimental settings, such as laboratory environments, or by adopting well-crafted experimental designs (Shadish et al., 2002), including between-subject RCT (Boudreau & Lakhani, 2016), within-subject (repeated-measures) designs (Keselman et al., 2001), regression discontinuity designs (Angrist & Pischke, 2009), and factorial designs, whose suitability and performance may depend on contextual constraints. The choice between these designs therefore reflects the epistemic function of the experiment and the constraints imposed by the phenomenology under study, rather than a hierarchy of methods (Falk & Heckman, 2009; Shadish et al., 2002).

The notion of internal validity is often contrasted with the notion of *external validity*, which refers "*to what populations, settings, treatment variables, and measurement variables can this effect be generalized*" (Campbell and Stanley 1963, p.5). The literature distinguishes several dimensions of external validity and corresponding strategies to address them, namely treatments (through variation and manipulation of treatment components), contexts (through multi-site or multi-context designs), target populations (through sampling strategies and population reweighting), and outcomes (through the use of alternative or expanded outcome measures) (Egami & Hartman, 2023).

The notions of internal and external validity are often considered orthogonal (Weimann, 2015) in that they capture distinct capabilities across different empirical (observational, experimental) and theoretical (analytical, computational) modeling approaches. While experimental researchers perceive a trade-off between increasing the rigor of internal control and maintaining the generalizability of findings, *some recent literature is supporting a more integrated theory-driven framework* (Angrist & Pischke, 2009; Bareinboim & Pearl, 2016; Heckman, 2020; Pearl, 2009). These approaches utilize causal theories and structural modeling to better support both internal and external validity, tailoring the research design to the specific domain of study and its unique contextual conditions.

Within this framework, standard laboratory experiments are typically characterized by strong internal validity and limited external validity, which may foster the perception that laboratory experiments are generally poor at generalizing, particularly to ecologically valid contexts. This characterization echoes earlier methodological literature, specifically Levitt and List (2007, 2008) who « argue that lab experiments are a useful tool for generating qualitative insights but are not well-suited for obtaining deep structural parameter estimates». This limitation is attributed primarily to low stakes, the use of student samples and participants' awareness of being observed (Hawthorne effect) all of which contributing to the pinpointed lack of ecological validity. Falk and Heckman (2009) argue that differences between laboratory and field outcomes reflect changes in underlying causal conditions rather than a failure of laboratory experiments per se. Furthermore, some influential comparative studies, such as Herbst and Mas (2015), highlight that, for *specific mechanisms*, behavioral patterns identified in laboratory settings can generalize to real-world contexts.

Current research thus generally supports the suitability of lab experiments for causal structural investigation and their applicability to several research domains (Weimann, 2015), while acknowledging a potential general lack of external validity in terms of sample representativeness and ecological realism, except for specific mechanisms where generalization is more plausible (Brüggemann & Bizer, 2016).

These considerations are particularly relevant to the distinction previously established between evaluative (use-oriented) and epistemic (explanation-oriented) experimental research. This dichotomy reflects a key characteristic of innovation studies: a large body of experimental work is policy-oriented, focusing on "what works?", that is, the evaluative capacity of laboratory experiments to measure the effects of specific interventions on target outcomes, including also those addressed in the traditionally non-experimental innovation policy literature (Bravo-Biosca 2020). In the context of policy-oriented research, lab experiments may encounter the difficulty of replicating the magnitude and significance of mean treatment effects when moving from controlled environments to complex, large-scale implementations (e.g., Al-Ubaydli, List, & Suskind, 2023). As highlighted by Brüggemann and Bizer (2016), the experimental approach has been widely adopted in policy-oriented innovation research across the three main domains: incentives and remuneration schemes; Intellectual Property Rights (IPR) and regulatory frameworks; and a broad category of institutional designs encompassing environmental policies, organizational structures, and behavioral interventions (nudges).

Another relevant part of innovation research is mechanism-oriented. This literature aims at identifying "what is the problem", that is, the structural causal processes underlying learning, creativity, coordination, and technological change, rather than evaluating the effectiveness of specific interventions. While often not labeled as such, this tradition aligns closely with a structural interpretation of causality and external validity, where generalization concerns the transferability of mechanisms across contexts rather than the mere replication of treatment effects. Classical domains within this strand investigates, among others, underlying incentives and motivation such as crowding out and crowding in (Boudreau & Lakhani, 2016); knowledge spillovers and diffusion (Boudreau & Lakhani, 2016); exploration–exploitation trade-offs; (Brüggemann et al., 2016) coordination and collaboration dynamics; and the design of platforms, contests, and tournaments (Deck & Kimbrough, 2017).

Finally, we conclude by emphasizing the often neglected but important role of laboratory experiments as valuable pedagogical tools. Historically, in the hard sciences (e.g., physics and chemistry) and in engineering, laboratory experiments have been adopted as valid learning devices (Lowe, 2023). Educational programs—from high school to university—commonly introduce laboratory experiments in these domains as training opportunities to acquire competencies and hands-on experience, as well as to reflect on key methodological issues inherent to experimental practice. This approach has also become increasingly common in social science educational programs. Through direct engagement with experimental design, researchers are

confronted with fundamental theoretical and methodological challenges, such as replicability, models of causal inference, statistical estimation, power analysis, and effect size interpretation. Addressing these issues helps develop methodological competence and sensitivity, enabling researchers to better identify weaknesses and potential pitfalls in their own empirical work (Holt, 1999; McCabe & Olimpo, 2020).

HOW TO USE IT

Inspired by the four types of validity proposed by Shadish et al. (2002), namely statistical conclusion validity, internal validity, construct validity, and external validity, and by the FAQs (frequently asked questions) proposed by Angrist and Pischke (2009), that is, the four fundamental questions to be asked in any successful research project, we propose a framework for the context of experimental research.

We retain the first FAQ "*What is the causal relationship of interest?*". This step is fundamental because it reorients the researcher away from the early descriptive–associative stage shaped by individual intuitions toward the explicit identification of cause–effect relationships. In the context of policy-oriented research, this activity can initially be reduced to a simple two-variable framework (cause–effect) and to the identification of the actual variables, or suitable quantitative proxies, that capture the phenomenon of interest. In the context of mechanism-oriented research, the need for a more explicit representation of the underlying causal structure, including mediators and confounders, can be addressed by a useful practice at this stage, namely formalizing modeling assumptions through causal diagrams (Directed Acyclic Graphs; see Pearl, 2009). This step requires, for an efficient and successful practice, a deep dive into the existing literature to precisely identify, extend, replicate, or validate established causal relationships.

We also retain in our framework the second question: "*What is the ideal experiment to capture the causal effect of interest?*". Before focusing on the first experimental solution suggested by personal experience, feasibility constraints due to logistics or local arrangements, or cost considerations, researchers should conceive the ideal experiment as, at a minimum, the hypothetical benchmark to which actual designs can be approximated in successive stages. This activity engages researchers in a careful reflection on the fundamental components of an experimental design such as the units of analysis, the target population, the treatments, the outcome variables, and the causal contrast of interest. This stage is fundamental for beginning to assess the presence of FUQs (fundamentally unidentified questions; Angrist and Pischke, 2009)). Indeed, some forms of randomization or manipulation may not be feasible in laboratory settings due to ethical, technical, or institutional limitations. In this step, we also lay the foundations for internal and external validity, because we define the context and limitations for, respectively, the nature of the causal claim of interest and its potential for generalization.

We rephrase the third FAQ as "*What is the feasible laboratory experimental strategy, given practical constraints?*". Indeed, our exercise is supposing the adoption of data coming from a lab experiment as *identification strategy*. This third phase enables researchers to conclude the operationalization of treatments and outcomes within feasible laboratory settings, including the specific tasks or tests used, in such a way that they truly represent the abstract constructs they are intended to measure. In this phase, the laboratory functions as an identification technology for causal inference (*internal validity*) and for achieving *construct validity*. Construct validity is achieved through the quality of manipulation, which ensures that the intended treatment is systematically and precisely attributed, and through the quality of measurement, which ensures that the outcome variables validly and reliably capture the phenomena of interest. This phase is crucial *because « the rewards associated with being correct in identifying causal relationships can be high, and the costs of misidentification can be tremendous » (Shadish et al., 2002).* In this phase, all aspects of the experimental procedure should be carefully devised. These include decisions about the materials and the experimental setting, including the coding of the experimental protocol and the preparation of instructions to minimize experimental demand effects (Shadish et al., 2002), as well as procedures to guarantee participant engagement and sustained attention. Additional control mechanisms typically include blinding or partial deception when appropriate, attention and comprehension checks and primarily randomization (most often implemented in between-subject design) and counterbalancing (most often implemented in within-subject designs) procedures (Falk & Heckman, 2009; Shadish et al., 2002). As a brief methodological note, while within-subject designs (both randomized and quasi-experimental) typically enhance causal identification by controlling for unobserved heterogeneity, they rely on a strict assumption of temporal stability that is often violated. Indeed, the repeated exposure inherent in these designs can trigger learning effects and experimental demand. Consequently, while within-subject designs offer superior statistical power, they may compromise ecological validity, making between-subject designs a more conservative but more successful choice for innovation studies.

The fourth and final FAQ is '*What is your mode of statistical inference?*'. This phase, which is rather technical, largely coincides with what is referred to as *statistical conclusion validity*, namely whether the statistical evidence is sufficient to support conclusions about the presence, magnitude, and direction of causal effects. This part includes pre–data-collection statistical considerations, such as decisions about the sample to recruit, power analysis for determining sample size based on significance levels and target effect sizes, as well as post–data-collection analyses, including the choice of statistical tests, the number of tests performed, and related inferential decisions. These statistical conclusions also inform, and pave the way for, inquiries into *external validity*.

Finally, to conclude this exercise, it is worth highlighting that current experimental research increasingly requires either time-stamped and publicly accessible pre-registration documents (Nosek et al., 2018), made available through dedicated web platforms (e.g., osf.io, aspredicted.org), or submission to peer-reviewed journals adopting the format of pre-registered reports (Chambers & Tzavella, 2022). This practice consists in making experimental designs public prior to data collection in order to enhance transparency, credibility, and causal interpretability, notably by reducing p-hacking and hindsight bias. The four FAQs constitute an exercise that, if conducted in a timely manner and properly documented, provides the natural foundation for such pre-registration practices and registered reports.

We also briefly describe the modern laboratories environments in which social science lab experiments are conducted. Nowadays, these dedicated spaces may be physical, online, or fully virtual, and they provide the controlled conditions necessary for systematically observing behavior, implementing treatments, and ensuring replicability.

Standardly, a laboratory for social and human sciences experiments may consist of a single workstation for running and investigating individual decision-making activities, or up to 20 to 30 separate computerized stations (the standard size for economics labs) for running many participants in parallel in individual or group-based activities (see Figure 2).

Computerized stations allow researchers to collect a variety of behavioral responses, such as choices made via mouse or keyboard clicks, as well as reaction times. However, additional equipment can be used to observe and investigate cognitive and emotional states and processes by monitoring physiological activity, such as electrodermal activity and ECG, eye movements using eye-tracking bars or glasses, neural activity using EEG headsets, and facial expressions using front-facing cameras (standard or infrared) (see Figure 3).

Fig.2. An example of a workstation for experiments measuring physiological signals and collecting both facial expressions with frontal camera and eye-tracking bar.

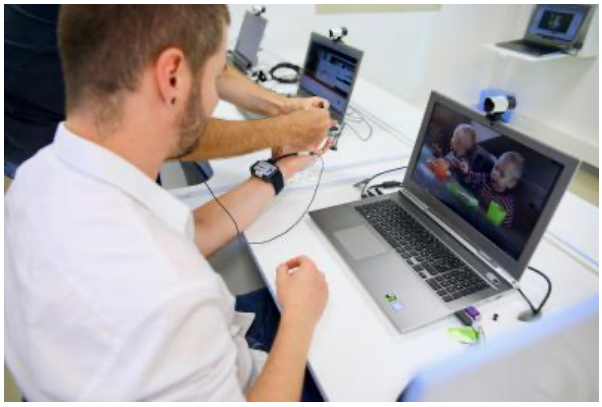

Fig.3. An example of a standard university laboratory configuration for conducting computerized experiments with multiple participants.

It is worth noting that the technical expertise and technological infrastructure required to run laboratory experiments in the social and human sciences have become increasingly accessible and affordable, largely due to free-to-use software solutions for implementing and managing computerized experiments (e.g., z-Tree, oTree) and for planning and administering randomized or constrained participant recruitment (e.g., ORSEE) (Fischbacher, 2007; Chen et al., 2016; Greiner, 2015). Many university laboratories offer experimental services such as support in the design and coding of the experimental protocol, as well as the independent running of experimental sessions, thereby providing turnkey solutions.

Furthermore, online platforms and solutions for experimental research have become increasingly adopted since the Covid-19 pandemic, now offering, at different levels, solutions enabling researchers to recruit hundreds or even thousands of participants (e.g., Prolific, Amazon Mechanical Turk), administer survey studies (e.g., Qualtrics, LimeSurvey, UserTesting and

Respondent), and run increasingly complex behavioural tasks (e.g., Gorilla, Pavlovia, Inquisit), all at reduced cost compared with setting up a new lab (Anwyl-Irvine et al., 2020; Bridges et al., 2020; Palan & Schitter, 2018; Peer et al., 2017).

CONFLICTS OF INTEREST

None to declare.

ANNOTATED REFERENCES

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2023). *The scale-up effect in early childhood and public policy: Why interventions lose impact at scale and what we can do about it.* Cambridge University Press.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Boudreau, K. J., & Lakhani, K. R. (2016). Innovation experiments: Researching technical advance, knowledge production, and the design of supporting institutions. *Innovation Policy and the Economy*, *16*(1), 135–167. https://doi.org/10.1086/684985

Bravo-Biosca, A. (2020). *Experimental innovation policy*. In B. H. Hall & N. Rosenberg (Eds.), *Innovation Policy and the Economy* (Vol. 20, pp. 191–232). University of Chicago Press/NBER. https://doi.org/10.1086/705644

Browder, R. E., Aldrich, H. E., & Bradley, S. W. (2019). The emergence of the maker movement: Implications for entrepreneurship research. *Journal of Business Venturing,* 34(3), 459–476. https://doi.org/10.1016/j.jbusvent.2018.10.005

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8, e9414. https://doi.org/10.7717/peerj.9414

Brüggemann, J., Crosetto, P., Meub, L., & Bizer, K. (2016). Intellectual property rights hinder sequential innovation: Experimental evidence. *Research Policy, 45*(10), 2054–2068. https://doi.org/10.1016/j.respol.2016.07.008

Brüggemann, J., & Bizer, K. (2016). Laboratory experiments in innovation research: A methodological overview and a review of the current literature. *Journal of Innovation and Entrepreneurship*, *5*, Article 24. https://doi.org/10.1186/s13731-016-0053-9

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, 6(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree: An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Deck, C., & Kimbrough, E. O. (2017). Experimenting with contests for experimentation. *Southern Economic Journal, 84*(2), 391–406. https://doi.org/10.1002/soej.12185

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, *8*, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

Gershenfeld, N. (2005). *FAB: The coming revolution on your desktop—from personal computers to personal fabrication.* Basic Books.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Hacking, I. (1983). Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge: Cambridge University Press.

Heckman, J. J. (2020). *Randomization and Social Policy Evaluation Revisited*. Institute for Fiscal Studies (IFS). https://ifs.org.uk/publications/randomization-and-social-policy-evaluation-revisited.

Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science*, *350*(6260), 545–549. https://doi.org/10.1126/science.aab0552

Holt, C. A. (1999). Teaching economics with classroom experiments. *Southern Economic Journal*, 65(3), 603–610. https://doi.org/10.2307/1061260

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, *54*(1), 1–20. https://doi.org/10.1348/000711001159357

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*(2), 153–174. https://doi.org/10.1257/jep.21.2.153

Levitt, S. D., & List, J. A. (2008). Homo economicus evolves. *Science*, *319*(5865), 909–910. https://doi.org/10.1126/science.1153640

Lowe, D. (2023). Rethinking the nature of experimental learning: Moving beyond conventional laboratory experiences. *European Society for Engineering Education* (SEFI). https://doi.org/10.21427/QW7S-N349

McCabe, T. M., & Olimpo, J. T. (2020). Advancing metacognitive practices in experimental design: A suite of worksheet-based activities to promote reflection and discourse in laboratory contexts. *Journal of Microbiology & Biology Education*, 21(1), 1–11. https://doi.org/10.1128/jmbe.v21i1.2009

Morris P. J. T. (2021). The history of chemical laboratories: a thematic approach. *Chemtexts*, *7*(3), 21

Nosek, B. A., et al. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*,

115(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114
Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004
Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595. https://doi.org/10.1214/14-STS486
Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Crowdsourcing participant pools for online experiments: A comparison of Amazon Mechanical Turk, CrowdFlower, and Prolific. *Behavior Research Methods,* 49(2), 433–442. https://doi.org/10.3758/s13428-016-0727-6
Pretzer, W. S. (Ed.). (1989). *Working at invention: Thomas A. Edison and the Menlo Park experience*. Johns Hopkins University Press.
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
Thomke, S. (1998) Managing Experimentation in the Design of New Products. *Management Science* 44(6): 743–762.
Weimann, J. (2015). The role of behavioral economics and experimental research in economics and policy advice. *Perspektiven der Wirtschaftspolitik*, *16*(3), 247–268. https://doi.org/10.1515/pwp-2015-0205