Philosophy of Science (2025), **92**, 470–487
doi:10.1017/psa.2024.23

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Re-Assessing the Experiment / Observation-Divide

Florian J. Boge

Institute for Philosophy and Political Science, Dortmund University, Dortmund, Germany
Email: florian-johannes.boge@tu-dortmund.de

(Received 15 September 2023; revised 12 February 2024; accepted 03 June 2024; first published online 05 December 2024)

Abstract

The article reevaluates the distinction between experiment and observation. It is first argued that to get clear on what role observation plays in the generation of scientific knowledge, we need to distinguish "experiential observation" as a concept closely connected to experience from "observation" in a technical sense and from "field observation" as a concept that reasonably contrasts with "experiment." It is then argued that observation construed as field observation can enjoy systematic epistemic advantages over experiment, contrary to appearances.

1. Introduction

Observations are central to empirical science, though what *counts* as an observation is all but obvious: van Fraassen (1980) coined a notion that allowed him to distinguish observation from inference by tying observation to unaided sense-perception; Shapere (1982) criticized empiricist notions as inappropriate to scientific usage, but his own account was criticized as too narrow (Bogen and Woodward 1988) or even off target (Linden 1992). So what *is* observation, and what *role* does it play in the generation of scientific knowledge?

Furthermore, there is a complicated relation between observation and *experiment* that "mainstream philosophy of science has had rather little to say about" (Okasha 2011, 223). On one hand, experiments seem unthinkable without observations: Michelson and Morley (1887) observed interference-fringes to determine earth's motion relative to the ether and Geiger and Marsden (1913) scintillations on a fluorescent screen to probe the nucleus's structure. On the other hand, "observational" is sometimes used as an *antonym* to "experimental," and we see claims to experiment's epistemic superiority over observation (Okasha 2011, 226–27; Woodward 2003b, 43–45). But this cannot be right if the preceding is too. So how *does* observation relate to experiment?

These questions can be answered only after due disambiguation. I shall hence distinguish "observation in the technical sense" (TO) from "experiential observation"

(EO) as a concept closely tied to experience and from "field observation" (FO) as a notion that reasonably contrasts with experiment.

This threefold distinction will prove helpful in answering questions concerning the epistemic role of observation in science. Specifically, I will here argue that FO is by no means generally epistemically inferior to experiment: in certain cases, it may even enjoy *systematic* epistemic advantages due to its unperturbing nature.

The first part (sections 2 and 3) introduces the three notions and their relations. This requires going into the relation between observations and data, as the kind of data taking distinguishes experiment from observation and is vital for evaluating the epistemic priority among them.[1] The second part (sections 4–6) then focuses on this epistemic priority among experiment and observation as recently scrutinized also by Boyd and Matthiessen (2023).

2. Three notions of observation

There have been various attempts to define "observation" in general terms, but all of them are wanting in some respect or another.[2] Arguably, this connects to the fact that scientists' use of "observation" "is typically relativized to the inquiry they have in hand" (Fodor 1984, 25). For instance, in high-energy physics (HEP), "observation" has a decidedly *statistical* character:

> if you want to claim, at least in high-energy physics, that you have observed a phenomenon, your result must be at least five standard deviations above background. (Franklin 2013, 1)

Thus "observation" here means an excess of specific activities in a particle detector that cannot be explained as a random fluctuation but indicates the presence of a sought-for particle.

In contrast, tissue biologists use advanced microscopes to gain insight into things like the interaction between nanoparticles and biological tissues (Jin, Bae, and Hong 2010). Atomic force microscopes, for example, direct a laser beam onto a cantilever with a sharp tip that interacts with the biological material through various forces. Because of this interaction, the cantilever is moved and the reflected light altered so that a differential image of the tissue is generated. In this way, biologists have "directly observed" the impact of nanoparticles on biological membranes through such things as the "formation of nanoscale holes . . . , membrane thinning, and/or membrane erosion" (Jin, Bae, and Hong 2010, 815).

[1] As a corollary, I will here shed light on the epistemic role of EO (which has traditionally been conflated with data taking) in instrument-heavy research.

[2] For example, van Fraassen (1980, 16) held that "seeing with the unaided eye" was "a clear case of observation," but this doesn't make contact with many scientific uses of "observation." Shapere (1982, 492) suggested that "observation" was information transmission without interference, trying to generalize from human vision, but this hardly fits some of the examples discussed here. Bird (2022, 159) recently suggested that an observation is a representation that causally depends on some fact about a given system and can serve as basic evidence, but it seems false that "the representation *is* the observation" (my emphasis).

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

These usages of "observation" in HEP and biology are clearly distinct. Yet they have commonalities: both involve close *causal contact* with the studied system (also Bird 2022). An "observation" in HEP is an observation because the relevant type of particle has been produced and decayed into characteristic products that *interact* with the detector so often that the resulting data cannot be discarded as a statistical fluctuation. Likewise, an "observation" in tissue biology is an observation because the cantilever interacts with the tissue through atomic forces.

A second unifying characteristic is that these are *success terms*: only if a certain level of statistical significance is exceeded, or an image *can* be interpreted as showing the action of certain nanoparticles, can observation be claimed. Hence I suggest collecting these different notions under a common header and speaking of "observation in the technical sense" (TO):

> *x* makes an *observation in a technical sense* (TO) on *y iff x* successfully establishes some relevant claim *c* about *y* by means of close causal contact with *y* within a scientific inquiry.

This defines a family of terms because standards of relevance and success vary with the field and context of inquiry, as the examples show. "Success" must not be misinterpreted though: null results can represent tremendous successes (think Michelson–Morley). What does forestall epistemic productivity is when research remains "inconclusive," that is, when the conditions for *applying* the respective notion of TO have been met neither positively nor negatively.[3]

"Observation" in a *nontechnical* sense is arguably different. "Seeing with the unaided eye" may be "a clear case" (van Fraassen 1980, 16), but only if this includes the *paying of attention* to a certain property, pattern, or object (Shapere 1982, 507). For instance, observing a bird in the backyard is distinguished from merely gazing out the window exactly by the fact that dedicated attention is being paid to the bird; observing the color shift of a TV is distinguished in the same way from watching TV. Hence I suggest introducing a second notion, which we may preliminarily define as the *paying of dedicated attention to an object of one's sense perception*.[4]

I claimed that all experiments involve observation, but as we saw, this is not true if we mean this in the sense of TO: some experiments are inconclusive and thus yield no observations in the technical sense. However, is it at least true that all experiments involve observation-as-perception-plus-attention? Bird (2022, 169–70) discusses a science fiction scenario in which knowledge from an experiment is fed directly into a subject's brain by means of an implant and so is gained without observation as perception-plus-attention. Hence there are *imaginable* experiments (and TOs) that could be done (or made) without perception.

[3] For example, the so-called 750 GeV bump in HEP was reported at a local significance of 3.9 standard deviations (ATLAS Collaboration 2016) and received tremendous attention but was later discarded as a statistical fluctuation.

[4] In science, this often involves the equipment as perception's object (Hacking 1983, 167), but "object" need not even be construed in any more involved sense (Chang 2005, 878): it could be merely a patch of color.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

This suggests that we should lean on a broader notion of "experience" than sense perception, for the subject would still *experience* the knowledge gain:

> *x* makes an *experiential observation* (EO) on *y iff y* is an object of *x*'s experience and *x* pays dedicated attention to *y*.

Now, if all experiments involve EOs and many even involve TOs, then neither EO nor TO defines a *contrast class* for "experiment." So how can we make sense of the distinction between experiment and observation indicated in the introduction? I suggest that we must acknowledge a third, distinct notion that thus sensibly contrasts with experiment: that of a *field observation*, which we may preliminarily define as the *unperturbed taking of data on an object of interest, that is, under natural conditions*. In contrast, "manipulation" and "control" are the key terms defining experimentation.[5]

For instance, consider how a team of biologists analyzing the correlation between vocalizations of male and female rhinoceroses and the testosterone levels in the males' feces during mating season with advanced software, technology, and statistics (Jenikejew et al. 2021) is seemingly engaged in a very similar activity as a team of particle physicists analyzing count rates of quantities computed from detector readouts with advanced software, technology, and statistics. However, while the biologists will take every precaution not to *disturb* the rhinoceroses, there is *no way* of measuring the relevant quantities pertaining to certain particles without exerting control over them.

Putting these intuitions into explicit definitions again would require discussion of further features of experimentation (such as repeatability; Currie and Levy 2019), but it will be sufficient to formulate *criteria* here that *partially* define both notions:

> Process *p* is a *field observation* (FO) of *y* by *x only if*, in the course of *p*, *x* takes data on *y* in an unperturbed fashion, that is, without *x* exerting control over *y* by relevantly manipulating *y*'s state.

In contrast,

> Process *p* is an *experiment* on *y* by *x only if*, in the course of *p*, *x* takes data on *y* while exerting control over *y* by relevantly manipulating *y*'s state.

These conditions naturally extend to collectives of scientists, when none/some of the scientists in the collective take data by manipulating *y*. They should be widely agreeable: Currie and Levy (2019, 1067, 1084) define experiments as "controlled manipulations" and contrast these with "observational fieldwork"; Boyd and Matthiessen (2023, 111) acknowledge a notion of "experiment as active manipulation," whereas "observation is … characteristically non-manipulative."[6]

[5] I say "manipulation" because "intervention" in at least Woodward's (2003b, 54) sense famously implies that one isolated causal connection is being probed, which is rarely the case.

[6] Similar intuitions are found already in Herschel (1830) or even Bacon (1620).

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

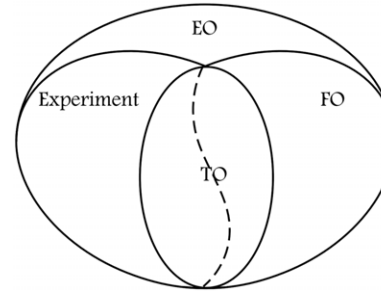

Figure 1. Relations between all notions of observation and experiment. The dashed line demarcates between TOs made as part of FOs or as part of experiments, respectively.

Earlier, I clarified the relation of EO and TO to experiment, but what is their relation to FO? Obviously, all FOs also involve EO and may generate TOs: a representation that provides evidence for certain phenomena may be generated or the statistical frequency of some type of event may exceed some threshold. In sum (figure 1), EO is the most encompassing notion, TO may occur as part of FO or experiment, and only FO contrasts with experiment.

3. Data taking and experiential observation

Prima facie, experimental control seems like a good thing: we can ensure (say) that particles collide where we want them to, in the quantities needed for TOs of Higgses. However, exertion of control means a perturbation of the studied system that may inevitably *destroy* subtle, sought-for effects. This issue will be centrally addressed later, but we should first clarify the notions of "data" and "data taking" centrally involved in the distinction between experiment and FO.

Empiricists like Hempel (1952, 21) famously put great emphasis on "data ... obtainable by direct experience," but in the age of complex experimentation and computer-aided data taking, assuming an intimate connection between data and EO seems inappropriate. This aspect is prominent in the work of Leonelli (2015, 812), who emphasizes that data are

> the results of complex processes of interaction between researchers and the world, which typically happen with the help of interfaces such as observational techniques, registration and measurement devices .... This is ... also the case for data generated outside the controlled environment of the laboratory.

Thus an ornithologist watching a bird needs to write down selective results from her EO, or use a digital camera to make images and video clips, to create data. These data then are "conditioned both by the employment of specific techniques and instruments ... and by the interests and position of the observer" (Leonelli 2015, 812).

Furthermore, in many scientific disciplines, even "raw" data are not connected to the experience of anything to do with the system under study. To draw on the example again, high-energy physicists call "raw" those data "arriving from an experiment's data acquisition system," which are then "organized in 'event records'" (Delfino 2020,

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

626): lists of numbers that constitute basic representations of the activity in the detector (see Jacobsen 2006, 4–5).[7] Data *taking* here takes place when measurable currents created by the interaction of "debris"[8] from scattering events with the detector arrive at the storage.

Indeed, "what counts as data," at least as *relevant* data, "depends on who uses them, how, and for which purposes" (Leonelli 2015, 811). For example, particles' energies, momenta, and angles relative to the colliding beams are usually computed as functions of HEP event records before analysis. Sometimes even higher functions are used, such as masses of decayed particles computed from energy-momentum-conservation, and these are simply considered "high(er)-level data."

So data are representations of systems' properties as exhibited in interactions, and raw data are generated by means of close causal contact. How and which of these properties are represented depends on the aims of the analysis.

Two things are noteworthy. First, in the preceding criteria for experiment and FO, I highlighted control and the unperturbing nature of the investigation, respectively. We can now make sense of this by taking into account the causal nature of data taking: if the act of data taking steers the studied system into a particular state, this cannot be an FO, though it might mean experimenting. If this feature is absent, this data taking cannot be part of an experiment, as control requires the manipulation of states. As I will argue, this can put FO at a *systematic advantage*, contrary to appearances.

Second, analyses aimed at establishing TOs target *data*, and we noted that EO is often quite distinct from data taking. Hence EO is not the main driving force behind the *inferences* made within those activities, so what *role* is left for it in modern science? I submit that EO usually functions as a *mediator* between FO or experiment and TO: *only* by witnessing certain displays on a computer screen, or by noticing information transmitted into the brain by a computer chip, can a scientist establish a claim of interest, based on experimental or field observational data.

As a corollary, empiricists remain at liberty to claim (van Fraassen 1980, 15) that our *interpretation* of EOs may change, leading to the reinterpretation and conceptual revision of many accepted TOs, but that the EOs themselves remain intact: EOs constitute the "phenomena" empiricists should *want to* save (Teller 2001, 135).[9]

4. Benefits of experimental control

A number of authors have addressed the question of why increased control over data taking, as involved in experimentation, might imply an epistemic advantage. I focus on two discernible claims to epistemic superiority: to an increased ability to establish causal dependencies and to an increased ability to confirm lawlike connections.

[7] Hence, if data cannot "be seen as straightforward representations" (Leonelli 2015, 811), the emphasis must lie on "straightforward," not "representation." Furthermore, representation is rarely straightforward and usually involves the kind of contextual features that Leonelli highlighted (e.g., Bailer-Jones 2009, 189ff.).

[8] Physicists speak of "final state products," whose etiology is complicated and theoretically only partly understood (Boge and Zeitnitz 2021).

[9] Such a position may not be far from views held by the early Carnap (1928), which still receive attention today (e.g., Leitgeb 2011; Chalmers 2012). Hence it requires further effort to show that empiricism verges on incoherence by making it "the aim of science to predict our perceptions of computer screens" (Bird 2022, 137).

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

The first claim has been voiced by many scientists (see Woodward 2003a, 88) and is an integral part of Woodward's own account of causation. Accordingly, the most valuable experiments are those that, like randomized controlled trials (RCTs), most closely approximate interventions.[10]

For instance, in medical RCTs (see Rothman, Greenland, and Lash 2008), patients are administered one of two treatments. The kind of treatment will be assigned at random, and one of them is typically a placebo, which can be safely assumed not to have the desired effect. Furthermore, randomizing eliminates the possibility of unconsciously selecting a group composition that by itself has an effect. In this way, many possible alternative causal chains from the initial conditions of the trial to the final outcome can be statistically nullified.[11] So a significantly better recovery in the treatment group suggests that the treatment has the desired effect.

The upshot is that experimental manipulations, interpreted as active changes in the causal variables describing a studied system's state, may offer a handle on seeing whether changes in X do cause changes in Y if they reasonably approximate "surgical" interventions, as other influences on Y have been (statistically, and approximately) eliminated. This level of control is clearly missing in FO: our ability to plausibly infer that X influences Y by means of FO may crucially depend on, say, the availability of different lines of sufficiently diverse evidence, and this availability may depend on pure happenstance.

Turn to the second claim of epistemic priority: that experiment increases our ability to confirm lawlike connections. A Bayesian argument to this effect has been given by Okasha (2011). Confirming a law $\forall x(Fx \rightarrow Gx)$ by an FO to the effect that $Fa \wedge Ga$ for some a can be problematic: in case the lawlike connection $\forall x(Fx \rightarrow Gx)$ doesn't make it any likelier to meet an F that is also G, conditioning one's credences on $Fa \wedge Ga$ won't increase the law's probability.

For example,[12] assume that, for some contingent reason, all meteoroids in our solar system happen to be such that meteorites landing on earth have diameter greater than five centimeters. Additionally, assume that a law ensures that meteorites on earth would end up being greater than five centimeters in diameter should these contingencies cease to exist. Does the law make it any likelier that the next meteorite will be greater than five centimeters in diameter? Given how the scenario was set up, this is doubtful.

In contrast, in an experiment in which all as are *prepared* to be F, Fa becomes part of the knowledge base, and the law is *bound* to receive confirmation from the observation that Ga, so long as $0 < P_{Fa}(Ga) < 1$—which should hold while we still seek confirmation. Thus, producing a small meteoroid and making it fall to earth, we would probably be able to observe a meteorite that *is smaller* than five centimeters.

[10] Recall that interventions produce changes in purported effect variable Y *only* via changes in purported cause variable X (Woodward 2003b, 54). However, Woodward is well aware that this condition can only be approximated, even in RCTs. Frisch (2014, 83ff.) offers an interpretation of these intuitions for the physical sciences.

[11] However, some randomization procedures do not result in "unconfoundedness" in the sense that the result of a treatment is independent of whether the treatment is actually given (Sävje 2021). Additionally, a treatment may admit multiple versions, which allows interference by confounders after randomization (Heiler and Knaus 2021, 7).

[12] I modify an example given by Okasha (2011, 227).

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

Naively read, this argument seems oversimplifying, because we *cannot* always prepare our *a*s to be *F*. This is certainly true in the meteoroid example, but that basically just says that the envisioned experiment is not feasible. However, we also cannot prepare the particles produced in proton–proton collisions to be Higgs bosons. Should we thus take it that we cannot experimentally confirm that Higgs bosons have a mass of approximately 125 GeV?

I believe this would mean overstating the argument's underlying intuition: even though our preparation method produces all kinds of things that are not Higgs events, we can at least *select for* such events by carefully selecting data points that fit the expected characteristics.[13] We then use these to confirm whether they exhibit a mass value expected on account of the "standard model." But of course, this is possible only because the conditions of proton–proton collisions at the Large Hadron Collider (LHC) in Geneva are well controlled and, *therefore*, well known.

5. Systematic benefits of field observation

As we saw, there *is* an epistemic benefit to experimentation in HEP, as the relevant information would likely be impossible to acquire under less controlled conditions. But this is just one example. *Generalized* claims to an epistemic benefit from increased control have often been embraced unquestioningly. Among the few to *argue* for control's benefits are Currie and Levy (2019, 1070ff.): according to them, control allows the isolation of a studied system from environmental factors so that one can *reproducibly* interact with the system's relevant properties and retrieve more fine-grained information for discriminating between hypotheses. However, whether an epistemic priority transpires from this *in general* still remains unclear.

In fact, Boyd and Matthiessen (2023, 123–66) have recently argued that it does not. In detail, Boyd and Matthiessen discuss the following factors that make an empirical activity epistemically privileged: signal clarity, characterization of backgrounds, and the discrimination and variability of precipitating conditions. Signal clarity means establishing the sensitivity of an apparatus to a given type of signal, as well as its being affected by processes not of interest, generically termed "noise" (124). "Backgrounds," in contrast, are data contaminations that "can be attributed more specifically to certain sources" (124). Finally, precipitating conditions are "the conditions that produce the signal in the first place" (125). Hence discriminating these means seeking out various causes of a TOed effect or signal.

Boyd and Matthiessen (2023) accomplish providing real-world examples whereby FOs can claim high performance on all these measures. This is an important achievement, but it does not quite establish whether *intrinsic* features of experiments can make them epistemically inferior (and hence FO intrinsically superior). In what follows, I discuss several cases in which the reasons for FO's epistemic superiority have to do with an intrinsic factor: the absence of control. I will coin these reasons "systematic," in contrast to "contingent" ones, where it just so happens that certain pieces of information can be obtained only by means of FO.

[13] This issue is complicated by the fact that quantum interference contributes a kind of "irreducible background" (Passon 2019; Schwartz 2021). But I will eschew the discussion of such quantum niceties here.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

To be clear on this issue, let me first briefly discuss those cases in which superiority does hinge on contingent factors. Astronomy provides a wealth of examples, as FOs here cannot be complemented by experiments (also Boyd and Matthiessen 2023). They are hence "all there is to go on" (Okasha 2011, 227). For instance, MIT describes the Even Horizon Telescope as "a group of observatories united to image the emission around supermassive black holes."[14] The use of "observatory" here reflects the fact that we cannot prepare black holes and investigate their properties in a controlled fashion, as it just so happens that human beings lack the relevant measures of size and energy to perform these experiments.

However, to gather strong evidence about the laws of relativity, we might *want to* experiment on black holes: this could give us an edge in finding deviations, thereby tentatively confirming certain approaches to quantum gravity.

A distraction might be created by cases in which there are systematic deficiencies to *actual* experiments, but an experiment that could ultimately overrule FOs seems feasible. An example is caffeine research, in which experiments and FOs tend to highlight conflicting aspects in relation to health: whereas FOs suggest health benefits, such as cardioprotective effects and decreased risk for development of type 2 diabetes or even neurodegenerative conditions, experiments suggest adverse effects, such as increased systolic and diastolic blood pressure or increased blood glucose levels (James 2018).

The main problem associated with the experimental evidence here is the *time scale*, for "acute physiological effects tend to . . . abate within hours," and RCTs have so far only been conducted on the scale of "weeks and months" (James 2018, 853). Hence there are limitations to the quality of experimental evidence that relates to intrinsic features of *actual* experiments but still falls short of establishing FOs' superiority in this case:

> Poorly understood confounder influence is a likely major cause of the enduring disjunction between the findings of experimental and observational studies . . . . Long-term randomised trials are needed to [understand] the health implications of lifelong coffee/caffeine consumption. (James 2018, 852–53)

In other words, the conflict between experimentation and FO here has nothing to do with features of experimentation per se: "coffee consumption is but one among numerous variables of life-style and environment," whence long-term experiments that control the "many factors" that "may confound the relatively weak coffee-health associations reported in the observational literature" (James 2018, 852) might settle the debate.

A similar distraction arises when surrogate systems are experimented on. Famous examples are analogue (Dardashti, Thébault, and Winsberg 2017) and "bottle" experiments (Currie 2020). Analogue experiments involve a system that is easier to handle than the system of interest but is assumed to share a set of common laws with it under specific conditions (Dardashti, Thébault, and Winsberg 2017, 63ff.). Dardashti, Thébault, and Winsberg and Dardashti et al. (2019) argue that this delivers a basis for confirming facts about the targeted system; others (e.g., Crowther, Linnemann, and

[14] https://www.haystack.mit.edu/astronomy/astronomy-projects/event-horizon-telescope.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

Wüthrich 2021) have been more skeptical. In any case, the fact that a different system is used makes this a *surrogate* experiment, something that has been suggested to define a general sense of *simulation* (Dardashti, Thébault, and Winsberg 2017; Boge 2019, 2020) or representation (Suárez 2004).

Due to the need for first establishing the connection between targeted system and system experimented on, it remains unclear whether such experiments are advantageous to FO if the latter is conducted on the right kind of system. But it seems clear that an experiment on the right kind of system *would be* advantageous.

Bottle experiments are another example (Currie 2020, 905), which, however, involves specimens from the relevant ontological domain. In ecology (where the term originates) these are experiments on "lab-raised, easily managed critters in highly artificial environments" (Currie 2020, 906). So, does this not provide an epistemic advantage over both analogue experiments and FO?

This seems doubtful, as the surrogate nature of bottle experiments nevertheless creates obstacles in confirming laws and causation, because it relies on what Currie (2020, 912) calls "extrapolationism":

> Surrogates, according to extrapolationism, target natural systems, and the resemblance between them facilitates extrapolating results from the former to the latter . . . . For the extrapolationist the value of an investigation is primarily due to its confirmatory prowess: it provides grounds for belief in some hypothesis pertaining to natural systems.

Despite the fact that a bottled ecosystem is an ecosystem, it is an additional assumption in need of justification that findings on the latter can be representative of those on its larger-scale counterpart.[15] Furthermore, these limitations are due to factors intrinsic to the experiment itself: they arise from the fact that a surrogate (scaled down or merely analogous) system is being used. However, as with the meteoroid case, this does not establish that an experiment on an entire ecosystem would not be advantageous over bottled experiment *and* FO.

None of these examples is thus convincing as an example of systematic advantages of FO. As a kind of proof of concept, note that Boyd and Matthiessen (2023, 120) discuss causal models by Spirtes, Glymour, and Scheines (2000) in which "observation can distinguish between two hypotheses that experiment cannot." Another such proof is delivered by the possibility of "intervention artifacts," as discussed by Craver and Dan-Cohen (2024, 259):

> Perhaps when I alters X it also influences the detection apparatus *via* a route that does not pass through Y. Or perhaps some intermediate variable S influences the detection in a way that foils our ability to assess the changes to Y.

However, we are here looking for *real-world* cases that exemplify systematic disadvantages to experimentation. Hence, to see the general kind of problem associated with experimental evidence at work, consider the so-called Hawthorne

[15] Currie (2020, 916) himself holds that bottle experiments can provide how-possibly understanding, which is clearly weaker than confirmation.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

effect (also Feest 2022). This effect was first discovered in experiments conducted by Roethlisberger and Dickson (1939) at Western Electric Company's Hawthorne plant that were supposed to investigate the relation between workplace illumination and productivity.

The findings were curious: "the illumination was decreased step by step," but "it was not until illumination in the experimental room was reduced to a level corresponding to moonlight that ... productivity finally started to decline" (Wickström and Bendix 2000, 363). Later analysis suggested that the detailed engagement with the workers, which was supposed to ensure their cooperation in the study, led to an increase in motivation, which fully compensated for the effects of decreased lighting. Thus the very act of making workers participate in the experiment was in large part responsible for the outcome.

Today, the "Hawthorne effect" is used as an umbrella term for any kind of effect whereby controlled data taking on human subjects influences their behavior, and the evidence for this is fairly robust (McCambridge, Witton, and Elbourne 2014). However, control is *definitive* of experiments. Thus, insofar as the data taking relevantly alters subjects' behaviors, an experiment cannot possibly reveal the sought-for information and enjoys a systematic disadvantage.

Now, data taking is involved in FO as well, and subjects might alter their behaviors in virtue of the very fact that data are being taken on them. Thus maybe there is no advantage to FO after all? This is indeed a problem, but there is the option of *concealing* the data taking in FO. By definition, this is not possible in experiment: its data-taking activities involve manipulating the investigated system's state.[16]

Concealment of data taking has been discussed in the marketing sciences as a means of compensating for Hawthorne-like effects (Grove and Fisk 1992). An example is "mystery shopping," whereby a participant (experiential) observer acts as a regular customer so as to not be recognized as an observer. Of course, this concealment might not work: the EOed subjects might notice some odd behavior from the participant observer or equally notice a hidden camera. But when executed skillfully, concealed FO can compensate for the problem of "fat-handed" manipulations, as involved in the Hawthorne effect.

An anonymous referee has confronted me with an interesting objection here: in so-called deception studies (Stricker 1967, 13), test subjects are misled about the attitudes, beliefs, and so on being probed. Hence, when the relevance condition involved in the partial definition of experiment offered earlier is taken into account, concealment of experimentation might be possible after all.

A prominent example is conformity experiments, such as those by Asch (1951). Here test subjects were instructed to offer perceptual judgments about sameness or difference between lengths of lines on paper. In reality, most participants were actors offering false judgments, and the conformity of actual test subjects' judgments to the majority was probed.

Deceptions like these might seem to mitigate Hawthorne-like effects. However, Schulman (1967, 27) early on demonstrated that subjects' responses varied as "a function of concern with the evaluation of [their] behavior," by varying "whether the experimenter and the group were perceived by the subject ... to observe (evaluate)

[16] This is noted by Boyd and Matthiessen (2023, 121), but they do not expound on the implications.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

[them]." In turn, this dependency might be mitigated by concealing the test subject from direct EO by other participants and the experimenter in the *response situation*. But regardless of this, participants' *suspicions* about the purposes of a given experiment remain a delicate matter: Stricker (1967) reported this issue to be underconsidered, inadequately probed (for example, by binarized variables), or underestimated in many psychological studies.

To date, methods for probing for suspicion are varied, as are estimates of the percentage of suspicious participants, and a unified framework is missing (Barrett, Neuberg, and Luce 2023). Furthermore, the use of deception methods within psychological experiments is now widely known, whence the worry quickly arose that participants would become more and more unreliable sources of information over time (Kelman 1967). Thus it remains a legitimate concern that the very act of making subjects participate in a study can distort their responses, and this sort of effect *cannot* be handled by deception.

This reasonably establishes that FO may be advantageous for certain purposes in psychological research, but does this issue pertain only to the social sciences? I believe the answer is no: an issue quite analogous to the Hawthorne effect can be straightforwardly seen to arise in natural science experiments, as preparing a physical, chemical, or biological system in a particular way may accidentally introduce additional effects that spoil the informativeness of the outcome.[17]

For example, Weber (2004, 287, emphasis omitted) points out that "preparation artifacts," which "arise when the biological specimen is fixed, cut, stained, or decorated for light or electron microscopy," are "one of the most frequent forms of error in biological laboratories." Thus, depending on the type of artifact, an experimental study of biological materials may well become uninformative about the properties investigated, and in virtue of the very preparation method. However, it remains unclear whether an FO could here yield the sought-for information instead.

A clearer example is provided by the conflict between internal and external validity in medical RCTs. "Internal validity" refers to a study's freedom from systematic biases, "external validity" to its generalizability. In RCTs, the attempt to achieve internal validity is "operationalized ... as inclusion and exclusion criteria," which lead to "a study population ... with increasingly controlled conditions" (Averitt et al. 2020, 1). However, a treatment might have a nonrandom variability across different subgroups (Varadhan and Seeger 2013), and this information can be lost by exclusion of relevant subjects.

So ensuring internal validity relies crucially on exerting control by handcrafting treatment and control groups. At the same time, this might spoil generalizability. In particular, one can apply eligibility criteria from RCTs to select data from an FO. If the RCT is externally valid, this should not lead to differences in the comparison between FO and RCT—but nevertheless sometimes does (see Averitt et al. 2020, 2ff.). This ostensibly shows that there are pieces of information (such as the influence of

[17] A naive reading of the quantum measurement problem has "observing" a quantum system "collapse its wave function," wherefore the very act of (experientially?) observing produces all outcomes of experiments on quantum systems. However, available interpretations of the formalism differ grossly, so this is more than controversial.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

"undocumented factors" on treatment variability; Averitt et al. 2020, 7) that are destroyed by the very act of exerting control.

I have provided two examples in which intrinsic disadvantages of experiment are salient and FOs exist that can arguably yield the sought-for information. What to conclude from this in general? The least we can say is that whether experiment or FO is advantageous is a case-by-case decision and that this is due to features that *make* an empirical inquiry an experiment or FO. However, I would also point out that it is usually very hard to tell what the overall effects of manipulation are. Hence, in disciplines ranging from physics to social science, researchers should value FO as a complementary source of information that need not be seen as generally inferior but can also provide hints as to where experiment might go wrong.

6. A strict dichotomy?

I proposed that data-taking activities that involve control over a studied system are not FOs, whereas those that don't are not experiments. This leaves it open whether there are data-taking activities that are neither. But are there any compelling cases?

Indeed, Perović (2021) argues that experiment and observation lie on a *continuum* but acknowledges that certain cases "are points at the far ends of the continuum in terms of their respective levels of manipulation." It is unclear to me whether the distinctions drawn above are not sufficient to cut that continuum in half.

First off, note the crucial qualifier "relevantly" in the criterion for FO. For instance, we may ask people to fill out a survey, and of course we would thereby manipulate their state, but not necessarily the *relevant* state: what the survey is supposed to find out is whether people antecedently happened to be in some state that led to certain responses in the survey. So carefully planned "observational studies" involving questionnaires may count as FO (rather than experiment) if they are indeed unperturbing in the desired sense.

Furthermore, consider the role of "field" in "field observation": experiments may famously also be conducted in the field (e.g., Morgan 2013), but this merely means that a system is studied, in a controlled way, within its natural *environment*. It doesn't mean that one leaves the system alone so that it exhibits its natural *behavior*. This, however, is what I take to be implied by the "field" in FO: that some naturally occurring sequence of states can be detected on y by means of data taking, without thereby running the risk of altering that sequence.[18]

In contrast, because "field experiments" are "experiments designed and carried out by scientists to ape ... laboratory conditions in the field" (Morgan 2013, 343), we immediately see that these are just specific experiments: "the interventions are controlled by means such as dividing subject units into treated and untreated groups in order that experimental effects can be isolated" (Morgan 2013, 343). The original Hawthorne studies may serve as an example exhibiting the *dis*advantages of experimentation *even in the field*.

Slightly more interesting are "natural experiments," which Woodward (2003b, 103) takes to be cases in which an intervention takes place without human action. As my account of experimentation decidedly involves *human* action, these still fall under FO,

[18] So carefully studying scientists' behavior in the lab could very well count as a sociological FO.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

while underscoring that FO can be epistemically equivalent or even superior to experiment. This is consistent with verdicts by Anderl (2016, 661), who describes them as "the direct equivalent of randomized controlled experiments in an observational situation," or Currie and Levy (2019, 1086), who hold that "there are significant analogies between experiments simpliciter and natural experiments": analogy and equivalence can meaningfully obtain only between things that are in fact distinct.[19]

A final issue that deserves attention is the fact that Mättig (2021, 14455) has recently called the LHC, which I have called an experiment, "a hybrid of experimental practices and observation":

> The collisions of interest are primarily not those of protons, but of the quarks and gluons inside the proton. These can hardly be varied by targeted intervention . . . . What the LHC delivers is a huge range of different final states. The "properties of interest" are obtained by selecting certain types of events, comparable to surveys of galaxies by telescopes. In consequence, the material information obtained from the LHC is a mixture of targeted intervention and observation. (14432–33)

So, should we say that the LHC inextricably intertwines FO with manipulation? I doubt it. First, note the tremendous degree of control exerted by physicists over the colliding protons. For example, the angle at which beams of protons cross is dynamically fine-tuned in the order of 10^{-2} radians so as to yield the greatest number of interactions in the right places.[20] Furthermore, following quantum field theory, particles like Higgs's are literally brought into existence in proton–proton scattering. If this doesn't count as "control" over relevant specimens, what does?

Of course, physicists *are* interested primarily in the interactions between quarks and gluons, not protons. Yet, it is fairly common that the targeted system can be controlled only indirectly: in vivo studies of the effects of drugs on an organ, say, will inevitably involve manipulating the entire organism. Nevertheless, such studies are straightforwardly considered experiments.

Finally, that properties of interest are "obtained by selecting certain types of events" is also rather typical for experiments. In particular, consider how the LHC serves *multiple* purposes: although it was designed primarily to search for the Higgs, it also serves the purpose of precision measurements on known particles and searches for new physics. Hence the "properties of interest" relative to one purpose define "background events" relative to another. But this says nothing over and above the fact that any measurement activity will also produce "noise," next to the (final) states of interest.

There might be additional reasons to see the LHC as an FO. For example, "heavy-ion collisions at the LHC recreate in laboratory conditions the plasma of quarks and

[19] Morgan (2013) refers to what was called "natural experiments" as "nature's experiments," whereas natural experiments for her involve "'reverse designing' the natural/social situation in its environment into an experimental one" (349). Despite her insistence on the contrary, I find it hard to see how this makes natural experiments in Morgan's sense not a special kind of field experimentation.

[20] See https://home.cern/news/news/accelerators/lhc-report-colliding-angle.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

gluons that is thought to have existed shortly after the Big Bang."[21] Thus, owing to the immense energies involved, the LHC can recreate "natural" conditions—conditions that have occurred *absent any human intervention*. And does that not make it FO by definition?

I believe concluding as much would be in error: just as FO can replicate experimental conditions when circumstances "happen to be" an intervention, some experiments can replicate natural conditions of interest. None of this speaks for a breakdown of a dichotomy between the two sorts of activities, with their complementary advantages.

7. Conclusion

I have argued that we need to distinguish between EO as dedicated attention to experience, TO as a family of technical success terms, and FO as the unperturbed taking of data. EO was argued to be distinct from data taking but to function as a mediator between an experiment or FO and its result in instrument-heavy fields. TO was argued to be that for which experiment and FO aim. Most importantly, FO was argued to be sometimes epistemically superior to experiment, and for systematic reasons: in some cases, the very act of exerting control forestalls the kind of TO that may be available in a carefully designed FO. Furthermore, because it is generally hard to estimate the overall effects of manipulating a targeted system, researchers might want to value FO as a complementary source of information not prone to the same kinds of error.

Acknowledgments. The research for this article was generously funded by the German Research Foundation (DFG) as part of the research unit "The Epistemology of the Large Hadron Collider" (grant FOR 2063) and my Emmy Noether Group ("UDNN: Scientific Understanding and Deep Neural Networks," grant 508844757). I have profited from comments by three anonymous referees, from multiple discussions within the DFG research unit "The Epistemology of the Large Hadron Collider," and from an internal conference on the experiment–observation dichotomy.

References

Anderl, S. (2016). "Astronomy and astrophysics." In P. Humphreys (Ed.), *The Oxford handbook of philosophy of science*, pp. 652–70. Oxford University Press.

Asch, S. E. (1951). "Effects of group pressure upon the modification and distortion of judgments." In H. Guetzkow (Ed.), *Groups, leadership and men; research in human relations*, pp. 177–90. Pittsburgh: Carnegie Press.

ATLAS collaboration (2016). "Search for resonances in diphoton events at tev with the ATLAS detector." *Journal of High Energy Physics* 2016(9), 1–50. https://doi.org/10.1007/JHEP09(2016)001.

Averitt, A. J., C. Weng, P. Ryan, and A. Perotte (2020). "Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations." *NPJ Digital Medicine* 3(1), 67. https://doi.org/10.1038/s41746-020-0277-8.

Bacon, F. (1620). *Novum Organum*. London: John Bill.

Bailer-Jones, D. M. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh: University of Pittsburgh Press.

Barrett, D. W., S. L. Neuberg, and C. Luce (2023). "Suspicion about suspicion probes: Ways forward." *Perspectives on Psychological Science* 0(0). https://doi.org/10.1177/17456916231195855.

Bird, A. (2022). *Knowing Science*. Oxford, New York: Oxford University Press.

[21] https://home.cern/news/series/lhc-physics-ten/recreating-big-bang-matter-earth.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

Boge, F. J. (2019). "Why computer simulations are not inferences, and in what sense they are experiments." *European Journal for Philosophy of Science* 9, 1–30. https://doi.org/10.1007/s13194-018-0239-z.

Boge, F. J. (2020). "How to infer explanations from computer simulations." *Studies in History and Philosophy of Science*, 82, 25–33. https://doi.org/10.1016/j.shpsa.2019.12.003

Boge, F. J. and C. Zeitnitz (2021). "Polycratic hierarchies and networks: what simulation-modeling at the lhc can teach us about the epistemology of simulation." *Synthese* 199(1-2), 445–480. https://doi.org/10.1007/s11229-020-02667-3.

Bogen, J. and J. Woodward (1988). "Saving the phenomena." *The Philosophical Review* XCVII(3), 303–52. https://doi.org/10.2307/2185445.

Boyd, N. M. and D. Matthiessen (2023). "Observations, experiments, and arguments for epistemic superi- ority in scientific methodology." *Philosophy of Science* 91(1), 111–31. https://doi.org/10.1017/psa.2023.101.

Carnap, R. (1928). *Der logische Aufbau der Welt*. Berlin: Weltkreis-Verlag.

Chalmers, D. J. (2012). *Constructing the World*. Oxford, New York: Oxford University Press.

Chang, H. (2005). "A case for old-fashioned observability, and a reconstructed constructive empiricism." *Philosophy of Science* 72(5), 876–87. https://doi.org/10.1086/508116.

Craver, C. F. and T. Dan-Cohen (2024). "Experimental artefacts." *The British Journal for the Philosophy of Science* 75(1), 253–274. https://doi.org/10.1086/715202.

Crowther, K., N. S. Linnemann, and C. Wüthrich (2021). "What we cannot learn from analogue experiments." *Synthese* 198, 3701–26. https://doi.org/10.1007/s11229-019-02190-0.

Currie, A. (2020). "Bottled understanding: The role of lab work in ecology." *The British Journal for the Philosophy of Science* 71(3), 905–32. https://doi.org/10.1093/bjps/axy047.

Currie, A. and A. Levy (2019). "Why experiments matter." *Inquiry* 62(9-10), 1066–90. 10.1080/0020174x.2018.1533883.

Dardashti, R., S. Hartmann, K. Thébault, and E. Winsberg (2019). "Hawking radiation and analogue experiments: A bayesian analysis." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 67, 1–11. https://doi.org/10.1016/j.shpsb.2019.04.004.

Dardashti, R., K. P. Y. Thébault, and E. Winsberg (2017). "Confirmation via analogue simulation: What dumb holes could tell us about gravity." *The British Journal for the Philosophy of Science* 68(1), 55–89. https://doi.org/10.1093/bjps/axv010.

Delfino, M. (2020). "Distributed computing." In C. W. Fabjan and H. Schopper (Eds.), *Particle Physics Reference Library, Volume 2*, pp. 613–44. Cham: Springer. https://doi.org/10.1007/978-3-030-35318-614.

Feest, U. (2022). "Data quality, experimental artifacts, and the reactivity of the psychological subject matter." *European Journal for Philosophy of Science* 12, 13. https://doi.org/10.1007/s13194-021-00443-9.

Fodor, J. (1984). "Observation reconsidered." *Philosophy of Science* 51(1), 23–43. https://doi.org/10.1086/289162.

Franklin, A. (2013). *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*. Pittsburgh: University of Pittsburgh Press.

Frisch, M. (2014). *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Geiger, H. and E. Marsden (1913). "The laws of deflexion of α particles through large angles." *Philosophical Magazine* 25(148), 604–23. https://doi.org/10.1080/14786440408634197.

Grove, S. J. and R. P. Fisk (1992). "Observational data collection methods for services marketing: An overview." *Journal of the Academy of Marketing Science* 20, 217–24. https://doi.org/10.1007/BF02723408.

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

Heiler, P. and M. C. Knaus (2021). "Effect or treatment heterogeneity? policy evaluation with aggregated and disaggregated treatments." *ArXiv*. https://doi.org/10.48550/arXiv.2110.01427.

Hempel, C. G. (1952). *Fundamentals of Concept Formation in Empirical Science*. Chicago: University of Chicago Press.

Herschel, J. (1830). *A Preliminary Discourse on the Study of Natural Philosophy*. London: Longman and J. Taylor.

Jacobsen, R. (2006). "From raw data to physics results, lecture 1." In HR-RFA (Ed.), *CERN Summer Student Lecture Programme Course*. Online: https://indico.cern.ch/event/430543/attachments/930594/1317911/ Lecture1.pdf (accessed 09/23).

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

James, J. E. (2018). "Are coffee's alleged health protective effects real or artifact? the enduring disjunction between relevant experimental and observational evidence." *Journal of Psychopharmacology* 32(8), 850–54. https://doi.org/10.1177/0269881118771780.

Jenikejew, J., J. Wauters, M. Dehnhard, and M. Scheumann (2021). "The female effect: How female receptivity influences faecal testosterone metabolite levels, socio-positive behaviour and vocalization in male southern white rhinoceroses." *Conservation Physiology* 9(1), coab026. https://doi.org/10.1093/conphys/coab026.

Jin, S.-E., J. W. Bae, and S. Hong (2010). "Multiscale observation of biological interactions of nanocarriers: From nano to macro." *Microscopy Research and Technique* 73(9), 813–23. https://doi.org/10.1002/jemt.20847.

Kelman, H. C. (1967). "Human use of human subjects: the problem of deception in social psychological experiments." *Psychological Bulletin* 67(1), 1–11. https://doi.org/10.1037/h00240720.

Leitgeb, H. (2011). "New life for Carnap's *Aufbau*? *Synthese* 180(2), 265–299. https://doi.org/10.1007/s11229-009-9605-x.

Leonelli, S. (2015). "What counts as scientific data? a relational framework." *Philosophy of Science* 82(5), 810–21. https://doi.org/10.1086/684083.

Linden, T. (1992). "Shapere on observation." *Philosophy of Science* 59(2), 293–299. https://doi.org/10.1086/289669.

Mättig, P. (2021). "Trustworthy simulations and their epistemic hierarchy." *Synthese* 199(5-6), 14427–14458. https://doi.org/10.1007/s11229-021-03428-6.

McCambridge, J., J. Witton, and D. R. Elbourne (2014). "Systematic review of the hawthorne effect: new concepts are needed to study research participation effects." *Journal of Clinical Epidemiology* 67(3), 267–77. https://doi.org/10.1016/j.jclinepi.2013.08.015.

Michelson, A. A. and E. W. Morley (1887). "On the relative motion of the earth and the luminiferous ether." *American Journal of Science* 34, 333–45. https://doi.org/10.2475/ajs.s3-34.203.333.

Morgan, M. S. (2013). "Nature's experiments and natural experiments in the social sciences." *Philosophy of the Social Sciences* 43(3), 341–357. https://doi.org/10.1177/0048393113489100.

Okasha, S. (2011). "Experiment, observation and the confirmation of laws." *Analysis* 71(2), 222–32. https://doi.org/10.1093/analys/anr014.

Passon, O. (2019). "On the interpretation of Feynman diagrams, or, did the LHC experiments observe h γγ?" *European Journal for Philosophy of Science* 9(2), 20. https://doi.org/10.1007/s13194-018-0245-1.

Perović, S. (2021). "Observation, experiment, and scientific practice." *International Studies in the Philosophy of Science* 34(1), 1–20. https://doi.org/10.1080/02698595.2021.1978038.

Roethlisberger, F. J. and W. J. Dickson (1939). *Management and the Worker*. Cambridge, MA: Harvard University Press.

Rothman, K. J., S. Greenland, and T. L. Lash (2008). *Modern Epidemiology* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Sävje, F. (2021). "Randomization does not imply unconfoundedness." *ArXiv*. https://doi.org/10.48550/arXiv.2107.141970.

Schulman, G. I. (1967). "Asch conformity studies: Conformity to the experimenter and/or to the group? *Sociometry* 30(1), 26–40. https://doi.org/10.2307/2786436.

Schwartz, M. D. (2021). "Modern machine learning and particle physics." *Harvard Data Science Review* 3(2), 14. https://doi.org/10.1162/99608f92.beeb1183.

Shapere, D. (1982). "The concept of observation in science and philosophy." *Philosophy of Science* 49(4), 485–525. https://doi.org/10.1086/289075.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press.

Stricker, L. J. (1967). "The true deceiver." *Psychological Bulletin* 68(1), 13. https://doi.org/10.1037/h0024698.

Suárez, M. (2004). "An inferential conception of scientific representation." *Philosophy of Science* 71(5), 767–79. https://doi.org/10.1086/421415.

Teller, P. (2001). "Whither constructive empiricism?" *Philosophical Studies* 106(1/2), 123–150. https://doi.org/10.1023/A:1013170506726.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon Press.

https://doi.org/10.1017/psa.2024.23 Published online by Cambridge University Press

Varadhan, R. and J. Seeger (2013). "Estimation and reporting of heterogeneity of treatment effects." In P. Velentgas et al. (Eds.), *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Rockville (MD): Agency for Healthcare Research and Quality.

Weber, M. (2004). *Philosophy of Experimental Biology*. Cambridge Studies in Philosophy and Biology. Cambridge University Press. https://doi.org/10.1017/CBO9780511498596.

Wickström, G. and T. Bendix (2000). "The "Hawthorne effect"—what did the original Hawthorne studies actually show?" *Scandinavian Journal of Work, Environment & Health* 26(4), 363–67. https://doi.org/10.5271/sjweh.555.

Woodward, J. (2003a). "Experimentation, causal inference, and instrumental realism." In H. Radder (Ed.), *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press.

Woodward, J. (2003b). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Cite this article: Boge, Florian J. 2025. "Re-Assessing the Experiment / Observation-Divide." *Philosophy of Science* 92 (2):470–487. https://doi.org/10.1017/psa.2024.23