European Journal for Philosophy of Science (2022) 12:13
https://doi.org/10.1007/s13194-021-00443-9

PAPER IN THE PHILOSOPHY OF THE SCIENCES OF MIND AND BRAIN

Data quality, experimental artifacts, and the reactivity of the psychological subject matter

Uljana Feest[1]

Received: 5 April 2021 / Accepted: 21 December 2021

Abstract

While the term "reactivity" has come to be associated with specific phenomena in the social sciences, having to do with subjects' awareness of being studied, this paper takes a broader stance on this concept. I argue that reactivity is a ubiquitous feature of the psychological subject matter and that this fact is a precondition of experimental research, while also posing potential problems for the experimenter. The latter are connected to the worry about distorted data and experimental artifacts. But what are experimental artifacts and what is the most productive way of dealing with them? In this paper, I approach these questions by exploring the ways in which experimenters in psychology simultaneously exploit and suppress the reactivity of their subject matter in order to produce experimental data that speak to the question or subject matter at hand. Highlighting the artificiality of experimental data, I raise (and answer) the question of what distinguishes a genuine experimental result from an experimental artifact. My analysis construes experimental results as the outcomes of inferences from the data that take material background assumptions as auxiliary premises. Artifacts occur when one or more of these background assumptions are false, such that the data do not reliably serve the purposes they were generated for. I conclude by laying out the ways in which my analysis of data quality is relevant to, and informed by, recent debates about the replicability of experimental results.

Keywords Reactivity · Experimental artifacts · Data reliability · Philosophy of data · Experimental inferences · Replication crisis · Philosophy of psychology

✉ Uljana Feest
feest@philos.uni-hannover.de

[1] Institut für Philosophie, Leibniz Universität Hannover, Im Moore 21, 30167 Hannover, Germany

Published online: 03 February 2022

1 Introduction

It is a fundamental feature of human beings and other animals that we react – sometimes unwittingly and sometimes on purpose - to our physical and social environments. When provoked, we often have angry reactions, when startled, we have a fear-reaction, when surprised with a gift we might have a joyful reaction. When exposed to a particular learning procedure in an experiment, our minds might react by forming new memories. When asked to answer to the questions on a memory test, we might react by interpreting the question and by following the instructions provided or by willfully ignoring them. We might also react by trying to guess what the purpose of the study is, which in turn might have an effect on our behavioral responses, etc.

Methodological writings in the social sciences (including psychology and economics) have long recognized a class of reactions peculiar to human beings, i.e., reactions that have to do with our knowledge of being part of a scientific investigation. For example, Rosenzweig (1933) pointed out that a crucial difference between experiments in (say) chemistry and psychology is that the units that are studied in psychology "have minds of their own," (Rosenzweig, 1933, 342), resulting in the possibility of "uncontrolled experimental materials, *viz.*, motives, [being] brought into the experiment" (Rosenzweig, 1933, 353). Some decades later, a version of this problem became prominent with the discovery of what has become known as the "Hawthorne" effect, i.e., the alleged effect of being part of a study, as opposed to the effects of a specific manipulation.[1] Similar issues have been addressed under the label of a "demand" effect, i.e., the effects of experimental subjects (consciously or unconsciously) trying to figure out what is being asked of them in the experiment and acting accordingly (e.g., Orne, 1962). Writing in the philosophy of economics, Jimenez-Buedo and Guala (2016) choose the term "reactivity" to describe the phenomenon underlying such effects: "[W]e shall call reactivity the phenomenon that occurs when individuals alter their behaviour because of the awareness of being studied" (Jimenez-Buedo & Guala, 2016, 11; see also Jimenez-Buedo, 2021).

Now, it would seem that the mere fact of research subjects altering their behavior when studied is not necessarily a problem, unless the altered behavior somehow gives rise to faulty data and/or distorted experimental results.[2] Within experimental science, faulty data and distorted experimental results are commonly described by invoking terms like "confounders" and "experimental artifacts." But what are the standards, relative to which specific experimental variables count as confounders? And what are the standards, relative to which data or results can be regarded as artifacts? These questions underscore my contention that in order to understand the notions of faulty data and distorted results, we need a philosophical analysis of what makes for "good" data and for "correct" experimental results.

[1] There has been some scholarly pushback on how pronounced this effect really was. (e.g., Jones, 1992).

[2] Jimenez-Buedo (2021) refers to this kind of reactivity as "malignant," contrasting it with "benign" reactivity.

In this article, I intend to shed light on these questions by providing an account of experimental artifacts as derivative of that of an experimental result. In doing so, I will be taking the notion of *reactivity* as a point of departure. Importantly, the understanding of reactivity invoked by me is broader than the one proposed by Jimenez-Buedo and Guala (2016) above, while subsuming their usage as a special case. I start out by noting that the term "reactivity" refers to a *disposition to react*. While it is surely the case that the awareness of being studied can give rise to specific kinds of pernicious reactivity, this is not the only kind of reactivity that plays a role in experiments. It is only by virtue of our disposition to react to stimuli and to experimental instructions that we can generate behavioral data at all. Nor is it the case that all experimental distortions are due to the awareness of being studied as there can also be other potentially confounding variables in experiments. Experimental subjects can react to features of the experimental situation in ways that are not intended by the experimenter. This highlights that (a) the reactivity (in my broader sense) of human subjects is a necessary condition for the production of experimental data, while (b) potentially giving rise to "faulty" data and, ultimately, to "distorted" results. The overarching contention of this paper is that the challenge of psychological experimentation consists in the construction and implementation of experimental designs *that simultaneously exploit and suppress the reactivity of the experimental subject*, with the aim of producing data that unambiguously speak to the object or phenomenon under investigation. The crucial challenge for experimenters, then, is to design the experiment in such a way as to tease out "the right kinds" of reactions, i.e., those that speak to the question or hypothesis under investigation, thus allowing for genuine experimental results. My aim in this paper is to provide an analysis of what this means.

In the following, I will begin by laying out a case study from recent psychology to illustrate my claim that psychological experiments involve the simultaneous creation and suppression of reactions, highlighting that this is achieved by a high degree of artificiality (Section 2). I will then (in Section 3) argue that this artificiality should not surprise us, given the well-known fact that scientific data are highly local and often don't bear any obvious resemblance to the phenomena under investigation (Bogen & Woodward, 1988). This will give rise to the question what considerations go into data production in the course of specific experiments. I will address this question by proposing a general account of the notion of an *experimental design*, which clarifies the rationale by which researchers take their experimental data to warrant inferences to specific experimental results. Section 4 will argue that my analysis of the notion of experimental design not only elucidates the notion of an *experimental result*, but also its complementary notion of an *experimental artifact*. The bottom line of my account will be that that the conclusion of an experimental inference can be regarded as a genuine experimental result if researchers are correct in their assumption that they have succeeded in the creation and suppression of all the relevant reactions, and, thus, that the data meet the standards required for the intended conclusion. Section 5 returns to the case study to consider the question how we should evaluate circumstances where experimental data are fickle and fragile, such that attempts at the simultaneous creation and suppression of reactions cannot easily be replicated. I will argue that this situation is likely to occur in cases

of considerable conceptual openness and I conclude with some considerations about exploratory research.

2 The creation and suppression of reactions: A case study

In 1988, a group of researchers (Strack et al., 1988) published an article in which they described two psychological experiments they had conducted in the field of emotion research. I will begin here by providing a rough outline of the first of these experiments[3]: Experimental subjects were told that they were participants in a pilot study to investigate how people cope when their dominant hand is impaired, and they have to substitute it with another body part. Each participant was handed a pen and they were divided into three groups a "lip group," "teeth group," and "non-dominant hand group." Subjects in the lip-group were told to hold the pen with their pursed lips, without the pencil touching their teeth. And subjects in the teeth group were asked to hold the pen with their teeth (see Fig. 1). Subjects in the non-dominant hand group held the pen in their non-dominant hand. Subjects were then asked to perform three tasks with the pens (such as connecting dots on a paper). Subsequently, all subjects were presented with some cartoons and asked to rate their funniness. Unbeknownst to the subjects, these funniness-ratings were the intended data of the study. The authors found that the subjects who had held the pens with their teeth overall ranked the cartoons as slightly funnier than the members of the other group.

I have chosen this case because it nicely demonstrates (a) the artificiality of experimental designs in psychology and (b) the extent to which the simultaneous need to elicit and suppress reactions from the experimental subjects is central to such design considerations. To explain these points, let us consider the purpose of the study. The point of the experiment was to investigate the *facial feedback hypothesis* (Darwin, 1872), according to which the facial expressions we make when experiencing certain emotions influence the intensity with which the emotion is experienced. So, for example, the act of smiling when happy might make me even happier. More specifically, the experiment was supposed to differentially test two competing hypotheses of how facial feedback occurs: According to the first, our mood improves when we smile because we perceive the smile and infer that we must be happy, which in turn prompts us to feel happier. According to the second, the mechanism is a more direct one from the muscles to the feelings and doesn't involve a cognitive inference. The authors pursued this second (non-cognitive) hypothesis.

The rationale behind the experiment was that the instruction to hold a pen with one's teeth was going to elicit a muscle contraction that mimics that of smiling (creation of a reaction), while the explanation of the purpose of the study provided to the experimental subjects was designed to suppress any interpretation of that contraction as a smile (suppression of a reaction), thereby ensuring that the data spoke to

[3] We will return to the second study in Section 4.1 below to illustrate how researchers can introduce modifications to an experimental design in order to improve data quality.

the hypothesis that no cognitive inference is required for the facial feedback mechanism, rather than being potentially distorted by subjects' knowledge that the relevant feature of the experiment was the "smile." In turn, the ranking of the feelings evoked by the cartoons was supposed to provide measurements of subjects' experienced emotions, thereby allowing researchers to test the hypothesis that a specific position of facial muscles (those typically activated when smiling) can increase the intensity of a positive emotion. The expectation was that the measured emotions in the group of subjects in the "no-dominant-hand" group would be in between the two others as the facial muscles were neither stimulated nor suppressed in this condition.

Clearly, the experimenters created a highly artificial situation. The point of this artificiality was to elicit specific reactions while suppressing others, with the aim of producing uncontaminated data. In turn, the relevant understanding of "uncontaminated" appears to be that there is a clear causal link between the manipulation and the intended target variable ("smile" muscle) as well as between the target variable and the variable of interest ("experienced emotion") as measured by the funniness-rankings. When I speak of data "directly speaking to" a hypothesis in the following, I have in mind a direct, uncontaminated, causal link between the experimental manipulation, the intended object of the manipulation, the intended object of the measurement, and the data that result from the measurement.

The notion that experimental data are purposefully generated (by contrast with the more traditional notion of data being "given" to the observer), draws our attention to the existence of scientific agents as designing and executing experiments. To put it differently, the flipside of thinking of the psychological subject matter as *reactive* is to recognize that the psychological researcher is *active*.[4] Experimenters manufacture scientific data for specific purposes, following design considerations that are supposed to ensure that the data speak to the question or hypothesis under consideration. They do so by purposefully exploiting and suppressing the reactivity of their subject matter in those very specific circumstances, and they do so under specific, local historical and material conditions.

This calls for a closer analysis of the relationship between the locality and context-specificity of experimental data vis-à-vis the broader (and presumably context-transcendent) purposes they are generated for. More specifically, it calls for an analysis of the logic by which context-transcendent results are inferred from context-specific data. The question touches on what has come to be known under the label of "external validity."[5] The literature on the internal/external-validity distinction is concerned with two distinct scenarios, namely (1) situations where the question is whether the results of a specific policy intervention, often investigated by means of an RCT, can be extrapolated to a different context, and (2) situations where the question is what kinds of inferences can be drawn from experimental data that were

[4] Obviously, the presence of a scientific agent as designing experiments is not specific to the reactivity of *human* experimental subjects. We will retun to specific features of psychological experiments in the concluding section of this article.

[5] In my mind the canonical formulation of this problem within philosophy of science remains (Guala, 2005)

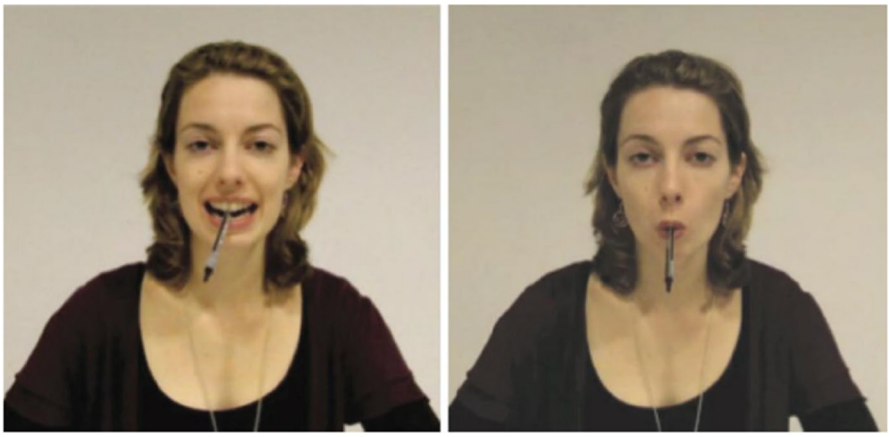

Fig. 1 Illustration of the instructions given to the test- and control group in Strack et al. (1988) (Source: Wagenmakers et al., 2016, 918)

generated to test a hypothesis about a general phenomenon assumed to exist outside the lab.[6] The difference is succinctly explained by Mook (1983), when he writes that in the latter scenario, "[w]e are not *making* generalizations, but *testing* them" (Mook, 1983, 378). For the purposes of this article, I am exclusively concerned with the latter of these two issues.[7]

In this vein, my question is not whether and how the result of an experimental inference can be extrapolated to a different context or generalized to a statement of a broader scope. Rather, my question is what kinds of design considerations go into an experiment, such that researchers feel warranted in using their data (which are produced locally and under artificial conditions) to make an inference about a hypothesis that is already pitched at a specific level of generality. Or, to put it differently, the question is in what ways the intended generality of the hypothesis under investigation has to be built into the experimental design. It is this question I turn to next.

3 Experimental designs as laying down conditions for experimental inferences

We begin with some brief considerations about the nature of experimental data as artificially and purposefully produced (3.1). After that, we will raise the question what kinds of design considerations need to be in place in order to treat experimentally produced data as allowing specific experimental inferences. This will prompt me to formulate a general schema of the logic of experimental design, crucially

[6] Jimenz-Buedo (2011) describes the former kinds of studies as "quasi-experimental field designs" and the latter as "pure lab experiments."

[7] I would like to thank Duygu Tunç for drawing my attention to the article by Mook.

highlighting, however, that experimental inferences rely on factual assumptions about the subject matter and its domain (3.2).

3.1 Manufacturing data, locally and idiosyncratically

The locality and artificiality of scientific data should not surprise us. In fact, Bogen and Woodward pointed to this feature of data a long time ago (Bogen & Woodward, 1988) when they distinguished between *phenomena* (the actual objects of scientific exploration and explanation) and *data* (the observable events that we use as evidence for claims about phenomena). As Bogen and Woodward famously noted:

> "Data are … idiosyncratic to particular experimental contexts, and typically cannot occur outside of those contexts. Indeed, the factors involved in the production of data will often be … disparate and numerous. … Phenomena, by contrast, are not idiosyncratic to specific experimental contexts. We expect phenomena to have stable, repeatable characteristics which will be detectable by means of a variety of different procedures" (Bogen & Woodward, 1988, 317).

Our case study illustrates Bogen and Woodward's point that data are highly contrived and carefully manufactured. They are designed to aid with the investigation of a (purported) phenomenon (in this case: facial feedback). They are not, themselves, the explananda of a scientific theory but are, rather, intended to serve as evidence for the existence of a particular phenomenon or for a particular hypothesis about a phenomenon. Applied to our case: There is nothing of inherent scientific interest about how people hold pens with their lips or teeth, or how they rank the funniness of a cartoon. Rather, the pen-holding serves the purpose of stimulating the muscles required for smiling, and the funniness-ranking serves the purpose of measuring the effects of the muscle-contraction on experienced emotion. Neither the experimental circumstances (the stimuli and instructions) nor the resulting data bear any obvious resemblance to the phenomenon under investigation. Nor is it clear that they should. Indeed, I argue that the necessity of experimental control highlights why artificiality is a crucial feature of experimental research. With this I do not mean to deny that artificially produced data can sometimes be misleading. However, I maintain that it is not their artificiality *as such* that can potentially create problems (see Feest, 2014; Jimenez-Buedo & Guala, 2016, for a similar point).[8] Problems arise if the artificiality fails to serve the intended purpose of creating and suppressing reactivity in the way required by the question under investigation. But what does that mean, exactly, and what are conditions of adequacy for artificially produced data?

In order to address these questions, I find it helpful to think in a principled manner about the various elements that need to be in place for a given inference from experimental data. I am inspired by an analysis of the elements of a research program, according to which we can distinguish between the "units" and "settings" of

[8] I return to this issue in Section 4.2 below.

experiments on the one hand, and the "treatments" and "observing operations" on the other (e.g., Cronbach, 1982). Importantly, an inference from experimental data to a specific conclusion is warranted if, and only if, all four of these units are chosen in a way that targets exactly the scope of the intended inference. In our example, the inference from the experimental data to a general claim about facial feedback is warranted only if (a) the experimental subjects (Cronbach's "units") are representative of the class of individuals intended to be covered by this claim (e.g., all adult humans), and (b) the experiment as a whole (Cronbach's "setting") adequately represents the spatio-temporal scope intended to be covered by the claim (e.g., all cultures at all times). Two additional requirements for a sound inference from the data to the correctness of the facial feedback hypothesis are (c) that the experimental manipulation (Cronbach's "treatment") causally affects precisely the variable that the claim is about (i.e., the facial muscle required for smiling), and (d) that the measurement instrument used to generate the data (Cronbach's "observing operation") measures precisely the variable thought to be affected by the muscle contraction (i.e., experienced emotion).

There is an important sense in which all four elements of a research design concern the scope of the inference we can ultimately make from the experimental data. This means that a comprehensive account of experimental inferences will have to cover all four. However, in this article my focus is on two elements that obviously bring up issues of *reactivity*: i.e., on the treatments and on the observing operations.[9] Consider, for example, the hypothetical cases where the teeth-treatment does not causally affect the variable of interest (the smile muscles) or where the data generated by the measurement-instrument are not causal effects of the variable of interest (experienced funniness). In such cases, we could say that the experimenter failed to produce the effects needed in order for the resulting data to speak to the hypothesis. Or, in the language I am using here, the researcher failed to adequately manage the *reactivity* of the subject matter. If either of these failures occur, the resulting data would be inadequate, either because the treatments and measurements fail to create any effect at all or because they trigger or fail to suppress an unwanted effect (possibly by accidentally triggering an unrelated variable or by failing to control for a variable that interferes with the intended effect). In addition, there might be confounders that occur (causally) between the treatment and the measurement, and which can occur even if the treatment- and measurement-operations as such are adequate. For example, the effect of the treatment on the measured variable might be distorted by the experimental subject's desire to please the experimenter.

To a first approximation, then, we may say that it is a necessary condition for the adequacy of experimental data that treatment and measurement in a specific experiment causally affect/detect the phenomena of interest and that there are no additional causal confounders. I believe that this analysis is compatible with recent works in the philosophy of data and evidence (e.g., Leonelli, 2015, 2020; Canali,

[9] I acknowledge that ultimately it is hard to disentangle reactivity from questions about the population and the spatio-temporal context since it is quite likely that reactivity is differentially moderated by these factors. We briefly touch on this issue in Section 5.2 below.

2020), which have pointed out that data quality needs to be evaluated relative to specific purposes. In the terminology developed here, if experimental data are created with the purpose of speaking unambiguously to a specific hypothesis, their adequacy has to be evaluated relative to this purpose. In psychology, this requires ensuring (among other things) that the reactivity of the subject matter has been managed adequately. As mentioned above, this needs to be evaluated in terms of all four elements of experiments mentioned above (i.e., not only with respect to treatment and measurement but also units and settings). Nonetheless, I think it is possible to bracket questions about experimental subjects and setting at least analytically and focus on what I call the "reactivity challenge."

3.2 The reactivity-challenge and the logic of experimental inference

As indicated above, the reactivity challenge consists in making sure that experiments only trigger the intended reactions while succeeding in suppressing unwanted reactivity. This amounts to making sure that experimental effects are not influenced by any unwanted causal factors in the experiment, i.e., by variables that might distort the processes required in order for the experimental data to serve as evidence for a specific claim. As we saw above, this can happen (a) if the manipulation causally affects something other than the intended variables (in our case: the relevant facial muscle), (b) the measurement procedure measures something than the intended variable (in our case: experienced emotion), or (c) there are any additional confounders that distort the causal process of the data-production. Given the above, we can break down the reactivity challenge into the "manipulation challenge" and the "measurement challenge," respectively (Danks & Eberhardt, 2009), as well as (what I will call) the "additional confounder challenge." These are the challenges of ensuring that the experimental manipulation does indeed causally affect (only) the variable of interest, that the experimental measurement does indeed (only) measure the variable presumed to be impacted by the variable, and (c) that there are no additional causal factors distorting the experimental effect, which might mislead us into misjudging the evidential status of a specific set of experimental data.[10]

It should be clear that it is no trivial matter to meet the challenges outlined above. Indeed, I would go so far as to claim that it is a defining feature of the epistemic situation of experimental psychologists that there is a great deal of epistemic uncertainty regarding the very subject matter under investigation, making it exceedingly difficult to figure out what one is manipulating and measuring, exactly, and which sources of reactivity to suppress (and how). Nonetheless, psychologists *do* draw inferences from their experimental data. So, the question is what is the underlying rationale of such inferences? How are experiments designed and implemented, and what kinds of assumptions need to be in place in order to regard any given inference as warranted?

[10] Notice that the confounder-challenge is not automatically taken care of if the manipulation-challenge and measurement-challenge are met since there could be confounders that are unrelated to the manipulation and measurement, including confounders that have to do with the choice of units and settings.

Experimental designs, on my analysis, encompass all the considerations researchers engage in to ensure that the resulting data allow for precisely the experimental inferences intended.[11] In other words, an experimental design aims to lay down the physical conditions that need to be in place, such that researchers are warranted in treating the data that result from the implementation of such a design as licensing an inference to a specific *experimental result*. An experimental result, then, is the conclusion of an inference from the experimental data.[12] This inference is warranted provided the experimental data stand in the right kind of causal relationship to the phenomenon under investigation. To return to our case study: The experimental result that facial feedback can be explained by a direct mechanism between facial muscle contraction and experienced emotion is warranted by the experimental data provided that (a) the treatment causally affected the relevant facial muscles and (b) the funniness ranking measures experienced emotion, and (c) there are no additional confounders. This amounts to saying that the inference to the experimental result is warranted, provided that the three above-mentioned challenges are met (the manipulation-challenge, the measurement challenge, and the "additional confounders"-challenge).

Let us think of an experimental result X as a proposition that is the conclusion of an inference of roughly the following form: "Phenomenon A (e.g., facial feedback) has feature B (e.g., being unmediated by interpretative processes) in population C (e.g., all humans) under circumstances D (e.g., everywhere at all times)." So, the question is what standards an experimental design and its implementation have to meet such that the data can be used to support this result.[13] As I indicated above, I am bracketing questions about population (units) and circumstances (settings). This leaves us with the question when a conclusion about a particular subject matter is supported by experimental data. I argue that one necessary condition is that the "reactivity challenge" is met, which breaks down into the three challenges outlined above. Put in the form of an inference schema:

P1: If I conduct manipulation A and get data B, then this implies result X, provided that the reactivity challenge is met, i.e.,

- P1a: manipulation challenge is met
- P1b: measurement challenge is met
- P1c: the additional-confounder challenge is met

[11] See Sullivan (2015) for a related analysis pertaining to cognitive neuroscience/neurobiology, and Lewis (2020) for an overview of the philosophical literature pertaining to experimental design more generally.

[12] My notion of an *experimental result* is similar to Sullivan's (2009, 2015) notion of an "interpretive claim."

[13] It is tempting to view this issue as related to that of internal validity as it has been conceptualized within philosophy of science (Guala, 2005). However, I would like to stay away from this terminology here, in part because I don't want to embrace the conceptual distinction between internal and external validity (see also other authors that have raised some concern about the coherence of the distinction, such as Mook, 1983, Jimenez-Buedo, 2011; Jimenez-Buedo & Miller, 2011, and more recently also Reiss, 2019). I will return to this in Section 4.2 below.

P2: conditions P1a-c are met.

P3: manipulation A is conducted and data B are observed.

C: result X.

Of course, this schema is basically that of a hypothetico-deductive inference. The experimental data support the result provided that a number of auxiliary hypotheses are true.[14] As Guala (2005, 109) puts it: Experimental inferences rely on a three-place relation: hypothesis, evidence, and background knowledge. What my analysis adds to this general point is that it provides a very specific account of what the relevant auxiliary hypotheses or background knowledge pertain to, namely to assumptions about what it takes to meet the manipulation challenge, the measurement challenge, and the confounder challenge *in any given case*. The flipside of making sure that these challenges are met is to probe for errors in the process of data generation (Mayo, 1996). My analysis suggests that in psychology such error probing involves taking seriously the reactivity of the subject matter and thus probing for unintended causal processes that might distort the data.

My analysis suggests something else as well, namely that finding out about possible errors can involve probing very specific factual assumptions about the reactivity of human (or animal) subjects. This means that even though probing for error importantly includes going through a general, or "trans-contextual" (Schickore, 2019, 212) repertoire of possible error sources (Mayo, 1996), it is not exhausted by them. For example, while certain potential confounders (such as demand effects) can occur in any psychological experiment, they may occur in specific ways, depending on the specifics of the phenomenon under investigation. Also, some confounders (such as confounders that might occur when trying to manipulate a specific facial muscle) only threaten a sub-set of experiments (i.e., those that manipulate a specific facial muscle). On the analysis presented here, probing for these latter, context-specific, kinds of errors is equally as important as probing for context-general ones. There is a slightly orthogonal point here: Mayo's context-transcendent error-repertoires are, as Schickore points out, "built with the benefit of hindsight" (Schickore, 2019, 212). By contrast, probing for context-specific errors can be part of the very process of gaining insight into the subject matter under investigation (Schickore, 2005).[15]

These considerations imply that even though my account suggests a somewhat formalized inference scheme, any individual inference crucially involves non-formal aspects that are highly specific to the subject matter at hand and to the local circumstances under which data are created. In this regard, the analysis presented here picks up on John Norton's contention that there can be no purely formal account of inductive inference (Norton, 2003). I follow Norton (2003) in referring to the additional, non-formal (factual) elements of empirical inferences as "material."

[14] As Tunç and Tunç (2020) point out, this means that experimenters face a situation of underdetermination.

[15] I am referring to this point as "slightly orthogonal" because, obviously, context-specific repertoires can be built in hindsight, too. However, Schickore's point, that error probing can have the function of not only ruling out confounders but also exploring the subject matter under investigation, is well taken. We will return to it in Section 5.2 below.

Summing up: Experimental inferences (inferences from experimental data to specific results) require highly specific material/factual assumptions, pertaining to the ways in which the reactivity of the subject matter can be stimulated and controlled. Experimental inferences are warranted if they are sound (i.e., if they are valid and the factual assumptions engrained in P2 are true). To come up with an experimental design is to create conditions that aim at making sure that the three components of the reactivity challenge are met. In turn, this means that if there are any doubts about specific experimental results, they are going to turn on the soundness of the inference from the data, and thus on the truth of the three assertions contained in premise 2.

4 (Experimental) artifacts and questions about data quality

The notion of an experimental result as being the conclusion of an inference from experimental data (provided the truth of P2) allows us to clarify the corresponding notion of an *experimental artifact* as pertaining either to questionable data-quality or unsound experimental inferences: Producing "good" data is (to a large part) a matter of managing the reactivity of the subject matter, i.e., eliciting and suppressing the right kinds of reactions. In this section, I elaborate on the question of data quality by arguing that my account of experimental artifacts is more comprehensive than existing philosophical analyses, while also capturing some of the insights articulated in them.

4.1 Experimental artifacts as (a result of) problematic data

Two instances of claims that somehow link experimental artifacts to problematic data can be found in Weber (2005) and in Guala (2000), where the former suggests that artifacts are *unreliable data*, and the latter suggests that artifacts are *mistaken interpretations of data*. Both of these suggestions speak to intuitions we might have about experimental artifacts, and both of them capture an aspect of the analysis I have presented above, namely that artifacts can occur when an experimenter draws a specific inference from experimental data, not recognizing that the data do not show what the experimenter thinks they show. However, neither of these accounts provide us with an account of how such mistaken inferences can come about. First, to grasp the notion of *unreliable* data we need a better understanding of what makes data *reliable*. Second, to grasp the notion of a *mistaken interpretation* of data, we need a better understanding of the logic by which experimental data are interpreted to begin with.

The analysis of experimental design presented in the previous section can shed light on both of these questions, in addition to showing how unreliable data and mistaken interpretations/inferences are connected. On my account, data are reliable if their causal history aligns with the presuppositions of an experimental design (i.e., if the assumptions built into P2 are in fact met, and, thus, that the reactivity of the subject matter has been exploited and suppressed in the right way). They are unreliable if they don't. Notice that the notion of *reliability* in play here is not the one used

in psychology, where the term simply refers to the *repeatability* of measurements. Rather, the notion is intended to describe data that are produced by "reliable" data production processes, which in turn means that the resulting data can legitimately be used as evidence for a given claim (Woodward, 2000; Sullivan, 2009). In the first (psychological) sense, an instrument can be reliable even if we don't know what it measures (i.e., it only has to reliably produce similar data). On the second, more demanding, understanding of "reliability," the data have to be produced by a procedure that meets the challenges laid out in P2 of the above inference scheme. This latter understanding is more demanding because it requires some assumptions (however preliminary) about the causal mechanisms (and possible confounders) involved in the data production. It is this later understanding that I endorse here.

Barring cases of fraud, I assume that researchers produce data with the intent of allowing for little ambiguity about the correct interpretation of (inferences from) the data. The correct interpretation of the data is supposed to be guaranteed by efforts to create data that speak directly to the hypothesis (i.e., to elicit and suppress reactions in "the right" way). Experimental artifacts occur when such efforts fail, such that the researchers mistakenly think that their data speak to the hypothesis, because they mistakenly believe to have ensured the truth of the auxiliary assumptions built into P2.[16] In other words: Experimental artifacts are the conclusions of unsound inferences, when researchers mistakenly think that they have reliable data. My analysis of experimental artifacts captures both the causal and the pragmatic aspect of Craver and Dan-Cohen (in press) recent "causal-pragmatic" account, in that (a) artifacts occur when something contaminates the causal structure that is required in order for data to legitimately play the role of evidence, and (b) the question of whether data are legitimate evidence depends on "what one is trying to measure or otherwise accomplish" (Craver & Dan-Cohen, in press, 22).[17]

My analysis of experimental artifacts/results sheds important light on investigative practices, specifically, practices that are designed to improve the experimental data relative to the question under investigation. As an example of this, let me return to the facial-feedback case study. As mentioned in Section 2, the authors of the original study (Strack et al., 1988) actually reported two very similar experiments, of which I have only described the first one above. The second experiment constituted slight variations of the first and was conducted because the authors were worried about the reliability of the data generated by the first. To put this differently, they were worried that the data of the first experiment did not license the inference they wanted to draw from it. For this reason, they introduced slight variations in their second study, "To strengthen the empirical basis of the results and to substantiate the validity of the methodology" (Strack et al., 1988, 772).

[16] Another way of putting this is to say that artifacts occur when experimental researchers fall prey to what Crasnow (2016) calls an "inferential bias," which can occur when their designs fail "to take into account factors that may have an influence on producing or preventing the production of an effect (also known as "confounders)" (Crasnow, 2016, 194).

[17] This latter point is similar to that of Canali (2020) mentioned above. I thank one reviewer for drawing my attention to Craver & Tal-Cohen's account.

Importantly, these variations were not random but addressed specific possible causal confounders in the first study. Specifically, they worried about two things: The first was that the experimental design of study 1 did not allow them to differentiate between the possibility that the muscle contraction *induced* the emotion as opposed to *modifying* an existing emotion (i.e., whether subjects thought the cartoon were funny because of the contraction of their facial muscle, or whether the contraction of the facial muscle merely had an impact on how funny they judged the cartoons to be). To differentiate between these possibilities, Strack et al., introduced two conditions in study 2, where in one the subjects had the pen in their mouths the entire time and in the other, they only had it in their mouth while doing the funniness rankings. This latter condition allowed them to rule out that the emotion had been caused by the muscle contraction (as opposed to merely being enhanced by it). The second worry was that the experimental design did not differentiate between (a) a cognitive judgment of the cartoons being funny and (b) the emotional experience of funniness. The researchers accordingly introduced two separate kinds of funniness rankings in order to ensure that their data spoke to the latter hypothesis.[18] The point I am getting at is that Strack et al.'s (1988) design considerations align with my analysis of how researchers should try to ensure data quality relative to a give intended inference.

My analysis can also account for a common usage of "experimental artifact" as describing cases where researchers literally manufacture the entity that they are intending to detect. A famous example of this kind of scenario is the case of the bacterial mesosome, where biochemists for some years thought that they had discovered a previously unknown cell organelle. While they did indeed detect a structure under their microscope, that structure later turned out to have been produced as an inadvertent side-effect of the techniques researchers used to prepare their samples for the microscope (see Rasmussen, 1993 for a detailed analysis of this case). While it is, of course, perfectly legitimate to say that the entity the researchers detected was an artifact of the method of preparation, this case can also be elucidated within the broader framework I propose here, i.e., as a problem of *data quality*. Researchers took their observational data (in this case, the data provided by their electron microscope) to license an inference to a specific result ("mesosomes are real"), but unbeknownst to them they had failed to meet the measurement challenge: their measurement outputs were caused by processes required for the application of the measurement instrument, and, thus, the researchers did not succeed in measuring what they thought they were measuring. The data produced by the measurement instrument were unreliable and did not license the inference the researchers had hoped to be able to make.

4.2 Ecological artificiality and higher-order artifacts

I have argued that the artificiality of experimental data is not problematic as such, provided that the data are reliable (in the sense outlined above), i.e., if they are the

[18] Guttinger (2019) uses the term "microreplication" to refer to such slight variations and makes the interesting argument that such microreplicatons are actually conducted much more frequently than recent worries about a replication crisis might suggest. I cannot judge the accuracy of this claim as a descriptive claim (though I have no reason to doubt it). But I certainly think it is accurate as a normative prescription on how to proceed.

result of procedures that have managed the reactivity of the subject-matter in "just the right way." Further, I have unpacked "just the right way" as meaning that the experiment meets the conditions laid down in P2 of our inference scheme.

But the question is whether this analysis really addresses the worry that experimental subjects are often manipulated in ways that create behaviors they would not exhibit under "real-world" normal conditions. This problem is articulated by Guttinger as follows: "When the entity or process of interest is placed in a context that is different from its native environment (in the case of biological entities or processes, this is usually the cell or the organism) there is a chance that behaviors are detected that are only specific to the new but not to the native context (or that native behaviors are completely lost)" (Guttinger, 2019, 461). In such cases, Guttinger remarks, we may be looking at artifacts.

In response, I would like to distinguish carefully between several problems. First (to repeat), the fact that organisms, when placed in artificial contexts, exhibit behaviors that they would not exhibit in their "native environments" does not automatically bar those behaviors from being good evidence for some claim about a phenomenon. My above analysis lays down minimal conditions they need to meet to be candidates for good evidence. However, I acknowledge that the objection still points to issues that need to be addressed. Two that I have in mind here arise from (1) cases where an experimental effect is indeed treated as evidence for the existence of a very similar effect outside the lab and (2) cases where the theoretical hypothesis that is being investigated by experimental means is impoverished or misguided vis-à-vis the reality it tries to capture, thereby giving rise to experimental tests that are equally impoverished or misguided.

Let's begin with the first of the two. In our case study it seems pretty clear that the experimental effect (i.e., the effect of holding a pen in a specific position on the ranking of the funniness of a cartoon) is not the phenomenon under investigation. There are other cases in psychological research, however, where researchers *do* wonder whether a specific causal effect in an experiment tells us something about the behavior of humans outside experiments. As Jimenez-Buedo and Guala (2016) recount, there is a debate about this in behavioral economics, which is concerned with precisely the question of whether some experimental effects are "caused by the peculiar circumstances in which the subjects are artificially placed" (Jimenez-Buedo & Guala, 2016, 5) rather than allowing us to learn something about a behavior that is also exhibited in the real world. The example they use concerns the dictator game, where experimental subjects are given a budget of 20 dollars and told that they can choose to transfer some of that money to an anonymous other player that has been randomly assigned to them. Surprisingly (from the perspective of utility maximization), many experimental subjects donate some money. This has given rise to the question of whether the effect is real (i.e., whether we can expect this kind of altruistic behavior outside the lab) or whether it is an experimental artifact. I argue that this question can be understood from within the framework of my analysis, in that it asks whether the experiment has introduced specific confounders that make the data unreliable with regarded to the intended conclusion about a phenomenon in the real world.

Let me now turn to the second way in which artificiality might be considered a problem. According to it, the problem lies not with the artificiality of the

experimental design or data as such (since, as I argued, what matters is that the data speak to the hypothesis), but with the lack of ecological plausibility of the *hypothesis* that is being tested. For example, take a research program about human development or human perception that focuses exclusively on cognitive mechanisms and ignores the sociohistorical and environmental circumstances in which human development takes place. From the perspective of more "ecological" approaches, this kind of hypothesis leaves out important variables.[19] An experiment that tries to translate such an impoverished hypothesis into an experimental design might be reliable in the sense defined above (i.e., speak directly to the hypothesis) but nonetheless end up producing experimental data that are not truly informative about the subject matter under investigation, because the hypothesis that is being tested represents an impoverished (or outright false) theoretical picture of the domain in question. In such cases, the artificiality-charge should be directed at the hypothesis or theory under investigation, not at the experiment.[20]

This possibility can be illuminated by means of Giora Hon's (1989) analysis of experimental errors. Hon distinguishes between four steps of an experiment, which include (1) laying down the theoretical framework, (2) designing and implementing an experiment, (3) making experimental observations and (4) processing/interpreting the data. Correspondingly, he says that errors can occur at each of these stages. While I have focused on steps 2-4 in this article, I want to highlight that the very choice of hypothesis and/or theoretical framework (Hon's step 1) might be misguided, such that experimental data, even if they speak to a hypothesis implied by this framework (i.e., are reliable in the sense outlined above), might be pointless from the perspective of a different framework or hypothesis. This analysis makes it clear that while artificiality is a desirable and necessary feature of stage 2 of an experiment, the resulting data are only epistemically valuable if – in addition to being reliable (i.e., addressing the hypothesis they are generated for) – the theoretical framework itself captures the complexity of the phenomena under investigation. In turn, this means that it is possible to generate "reliable" data for what may by others be regarded as an ecologically implausible theory.[21]

Hon's analysis is illuminating in that it reminds us that possible errors pertaining to experimental results are not exhausted by a failure to meet the reactivity challenge (in the sense outlined above) but can also be due to flaws at other points in

[19] Examples of this are Uri Bronffenbrenner in the realm of child development and James Gibson in the realm of perception.

[20] In this vein, it might be argued that ecological plausibility should be regarded as a theoretical virtue (see Tulodziecki, 2013 for a discussion of theoretical virtues).

[21] See Feest (2014) for an example from the history of psychology, having to do with atomistic vs. holistic approaches to perception. Atomists hypothesized that human perception can be reductively explained as being comprised of simple sensations that stand in a one-to-one relationship to simple stimuli, and they demonstrated the existence of such simple sensations by creating experimental circumstances under which they could be isolated. In turn, Gestalt psychologists did not deny that it was possible to isolate simple sensation by experimental means, but they doubted that this shed any light on the basic constituents of perception since we are rarely, if ever, exposed to simple sensory stimulation. Instead, they designed experiments to investigate the law-like ways in which we respond to complex stimulus-configurations.

the experiment.[22] For my current purposes, the "error" of testing an ecologically implausible theory is interesting. This kind of error is not strictly speaking an "experimental" error but it still results in problems with the inferences that can be drawn from the experimental data. In this vein, I distinguish between (1) experimental artifacts (unsound conclusions that are drawn from experimental data due to a failure to meet the reactivity challenge) and (2) higher-order artifacts (problematic conclusions that are drawn from experimental data due to the ecological implausibility of the hypothesis under investigation). I argue that higher-order artifacts can occur even if the reactivity challenge is met (i.e., if the data are reliable).

But there is also the converse question, i.e., whether it is possible to generate reliable data (thereby avoiding experimental artifacts) in cases where researchers aim to test an ecologically *plausible* hypothesis, i.e., a hypothesis that takes into account the complexity of the real world, or, as Guala (2003) puts it, experiments that are designed "with a concrete, specific target in mind" (Guala, 2003, 1204). This question obviously touches on the debate about the relationship between "internal" and "external" validity (see footnote 14 above), though I prefer to follow Sullivan's framing of the problem in terms of a tension between "reliability" and "validity" (Sullivan, 2009, 535). Sullivan argues that the requirements of reliability and validity pull in different directions, because the former calls for the simple and unambiguous experimental designs, whereas the latter demands that the complexity of the real world be brought into the experiment, "which would inevitably lead to a decrease in the simplicity of the effect produced in the laboratory" (Sullivan, 2009, 535). In turn, this would imply that with an increase of the complexity of an experiment it becomes harder to meet the reactivity challenge, i.e., to achieve reliable data that speak unambiguously to the hypothesis. Engaging with this important suggestion would require a more in-depth discussion of whether there is a necessary connection between the complexity of a "real-world" phenomenon and the complexity of hypotheses (and experiments) about the phenomenon. Such a discussion would go beyond the scope of this paper, however.

5 Reactivity, fragility, and replicability

On the account presented in this paper, experimental data are "good" if they succeed in supporting inferences to the intended conclusion. Achieving good data is contingent on meeting (among other things) the "reactivity-challenge," i.e., the challenge of ensuring that the experimental data are not distorted (relative to their purpose) by a failure to meet the manipulation, measurement, or other-confounder challenge. This raises the question of how data (and the inferences drawn from them) should be evaluated when we cannot be sure whether the background conditions required for a sound inference from the data hold. In this section I address this question by considering what replication failure might tell us about (a) about data quality on the one hand and (b) the "fragility" of the phenomenon on the other. I will argue that while robust experimental effects are necessary (though not sufficient) for experimental

[22] This is also remarked by Mayo (1996) and Sullivan (2015).

inferences, fragile effects can prompt investigations into the context-sensitivity of phenomena in the real world.

5.1 On the fragility of experimental effects

The facial feedback hypothesis study discussed above was recently subject to a large-scale replication study. In it, a group of researchers, with the help of one of the authors of the original study, devised a protocol that was closely oriented on Strack et al.'s (1988) study (Wagenmaker et al., 2016) and then ran the experiments in 17 different labs in Europe and the United States. They found that the initial effect could not be replicated.

Failure to replicate is never good news for the authors of a study. The fact that the 1988 facial feedback data could not be replicated should therefore certainly give us pause as it questions not only what was assumed to be a robust experimental effect but also the hypothesis the data of this study were supposed to provide evidence for. There have been a lot of debates recently about whether questionable research practices in the original studies (p-hacking, data massaging, retrospective hypothesis fitting) are to blame for replication failures (see Romero, 2019 for an overview). While I certainly do not intend to discount or downplay the problem of questionable research practices, I want to pursue a different question here: Given the reactivity and complexity of the subject matter as well as the degree of epistemic uncertainty regarding the auxiliary assumptions contained in P2, we should perhaps not be so surprised to find that many effects fail to replicate. In this vein, I suggest that we treat this as a problem of data quality and consider the possibility that either the original or the replication study failed to meet the reactivity challenge.

Assuming that there was no foul play involved in the 1988 study by Strack et al., one possible explanation for Wagenmaker et al.'s (2016) failure to replicate its experimental effect may have been that there was a subtle but profound difference between the original study and the replication study, which resulted in a failure to meet the reactivity challenge in one of them. And indeed, there is a recent study (Noah et al., 2018) that tests the hypothesis that Wagenmaker et al.'s (2016) failure to replicate Strack et al.'s experimental effect might be due to the fact that they introduced one crucial novel feature, namely telling subjects that they were being filmed. The authors re-did the original experiment under two conditions, one in which subjects were told that they were being filmed and one in which they were not told that they were being filmed. They were able to create Strack et al.'s experimental effect under the second, but not under the first condition (Noah et al., 2018). This suggests that Wagenbacher et al.'s failure to replicate the original effect (and thus be able to infer the same result) may well have been due to a classic reactivity effect (awareness of being observed).

In the current section, my concern is not with this particular case but with a fact it draws attention to, namely that the effects studied by psychologists are often "subtle and fragile" (Strack, 2017, 2). The notion of fragility is intriguing as it is in philosophy often contrasted with that of robustness, where the latter carries with it the connotation that perhaps the effect is not "real" (cf. Wimsatt, 1981). This does

not follow, however: An effect can be real, yet occur only under very specific circumstances. So, what do difficulties to replicate an effect tell us about the nature of experimental data and its relationship to the claims they are taken as evidence for?

To address this question, let me start by considering the notion of *fragility*. We can distinguish between the question of (a) whether the facial feedback phenomenon (in the real world) is subtle and fragile (i.e., highly context-sensitive) and the question of (b) whether the experimental effect is subtle and fragile and thus hard to create, which means that it is hard to gather evidence in support of the hypothesis that this phenomenon exists in the real world. The two can be, but need not be, related. A phenomenon, such as gravitational radiation, can be robust, but it can still be very hard to detect it experimentally. This has been famously (and notoriously) discussed in Harry Collins's work on the decade-long efforts to build a detector for gravitational waves, and to replicate existing data purporting to be picking up on gravitational waves (Collins, 1985, 2004).

In psychology, the possibility that experimental effects might be fragile is an obvious consequence of what I have referred to as the "reactivity" of the psychological subject matter. This means that the mere fact that an experimental effect is unstable (cannot be replicated) does not mean that the existence of a real phenomenon is thereby disproven. It may just mean that the subject matter is highly reactive to small changes in the experiment, which makes it hard to meet the reactivity challenge. Conversely, the analysis of data-quality (or data reliability) presented above highlights that replicability of experimental effects, while necessary, is certainly not sufficient to underwrite an inference from the data to a specific experimental result, since (as I argued above) the data also need to stand in the right kind of causal relationship to the events described by the experimental result. Thus, even if the experimental effect reported in the 1988 study had been reproduced in the 2016 replication study, this would not necessarily provide good evidence for the facial feedback hypothesis. In order for that to be the case, we would need to be certain that the intended inference from the experimental effect to the hypothesis is sound, i.e., that the manipulation challenge, the measurement challenge, and the confounder challenge have been met.

Summing up, while it seems clear that achieving a repeatable experimental effect is a necessary condition for the effect to serve as experimental evidence, this needs to be supplemented with the requirement of meeting reactivity-challenge (i.e., in order to serve as evidence for a claim about a particular phenomenon, it needs to stand in the right relationship to the facts described by the claim). But there is a follow-up question here: Given that the reactivity of the psychological subject matter can result in fragile experimental effects, what does this tell us about the phenomenon itself? If small changes in the experiment can make an effect disappear (as illustrated by Noah et al.'s analysis of the Wagenmacher et al., 2016 replication study), should we conclude that perhaps the phenomenon itself is fragile (in the sense of being highly context-dependent)? Above I have argued that the fragility of an experimental effect is compatible with the reality of the phenomenon it is evidence for (as in the case of gravitational radiation). But perhaps in psychological research it is the phenomena themselves that are fragile. It is this question we turn to next.

5.2 Fragile phenomena and the epistemology of exploration

There are two ways of construing the above question of whether the fragility and sensitivity of an experimental effect points to the phenomenon itself being fragile and context-sensitive: The first is that perhaps the purported phenomenon only occurs within the very specific context of the experiment that tries to generate evidence for "it." The second is that the experimental circumstances under which the evidence is generated are representative of some features of real-world circumstances under which the phenomenon manifests itself. Relatedly, perhaps the experimental circumstances under which the effect disappears (for example, when a camera is introduced into the experiment) might be representative of circumstances that moderate the phenomenon in the real world.

The former construal (where the phenomenon itself only occurs under the highly controlled circumstances of an experiment) raises the worry that we are looking at the kind of scenario described in footnote 21 above, where it was possible to generate reliable evidence for a hypothesis that was later deemed to be ecologically implausible. This might lead us to posit that fragility and context-dependence of experimental effects are indicative of there not being a stable phenomenon outside the lab.[23] The latter construal directly contradicts this pessimistic conclusion, suggesting instead that the context-sensitivity of the phenomenon within the lab is indicative of a context-sensitivity of the phenomenon outside the lab. In such a case, if we find a confounder (relative to a specific hypothesis about a phenomenon), this can give rise to a novel hypothesis about the phenomenon, namely that in the real world it is moderated by a specific variable. This is essentially the insight Noah et al. (2018) are drawing when they argue that "minute differences in the experimental protocol might lead to theoretically meaningful changes in the outcomes" (Noah et al., 2018, 657).

Psychologists frequently investigate the influence of moderator variables on their phenomena of interest (Hayes, 2014). Typical examples of moderator variables are age, gender, ethnicity, education, etc. (i.e., features of the experimental subjects). However, moderators can also be situational factors that are engrained in the setting of an experiment. Importantly, one and the same variable can be a confounder in one context (i.e., be something that needs to be controlled when pursuing a specific hypothesis) while being a moderator in a different context (i.e., something whose impact on the phenomenon can be investigated).

From the perspective of these considerations, then, the fragility of experimental data should not exclusively be viewed as raising questions about their reliability with respect to specific intended inference, but should also prompt us to formulate new hypotheses about the shape of the phenomenon under investigation and to test them by systematically varying the experimental conditions in accordance with specific hypotheses about possible moderators. We may refer to this kind of research as "exploratory" (Steinle, 1997), because it is aimed at exploring the shape of a phenomenon. However, it bears stressing that this kind of exploratory research should not be contrasted with research that is informed by theoretical hypotheses (Colaco, 2018).

[23] I am very grateful to one of the referees for helping me draw this connection to the discussion of ecological plausibility in the previous section.

On my analysis, an experiment can simultaneously probe into the reliability of the data generated by a previous experiment while at the same time furthering our understanding of the shape of a given phenomenon. The context-dependence of phenomena does not necessarily refute previous assumptions about their existence, though it will prompt us to narrow down the scope of our previous claims about them. I argue that any given experiment contains ambiguities that allow us to design follow-up experiments, which specifically address possible confounders and/or moderators, thereby allowing for a more nuanced understanding of the phenomenon (see also Feest, 2016, 2019; Boyle, 2021; Rubin, 2020). These ambiguities are a consequence of the uncertainty researchers face about their phenomena and the exact circumstances under which they occur.[24]

The viewpoint pushed by this kind of approach recognizes that our understanding of how data and phenomena are related can shift in the course of a dynamic research process, where our prior conceptualization of the phenomenon will shape the way we design an experiment, but the resulting data can also prompt us to reconceptualize our understanding of the relevant phenomenon in accordance, thus making conceptual progress (Feest, 2011; Potters, 2019). Relating this point specifically to the replication crisis in psychology, Lavelle (2020) argues that "failed replications can play a crucial epistemic role in allowing researchers to scope the limits of the effects under study, and to better understand the conditions under which they manifest" (Lavelle, 2020, 22). This can give rise to improved concepts of the relevant phenomena that inform, in an iterative manner, the designs of further experiments.[25] This way of proceeding also takes up Tal (2020) suggestion that descriptive research should play a bigger role in psychology.

6 Conclusion

In this article I have used an analysis of the notion of reactivity to shed light on a variety of issues in relation to questions about the quality of data and evidence in experimental research in psychology.

I argued that once we appreciate the reactivity of the subject matter in psychology, we can put into sharp relief the challenges that researchers need to tackle if they want to produce experimental data that speak to their hypothesis in an unambiguous fashion. After highlighting that experimental data production in support of a hypothesis has to both elicit and suppress reactions, I identified three challenges that need to be addressed by experimenters. I subsumed these three challenges under the label "reactivity challenge." Correspondingly, I highlighted that in order for experimental inferences from data to results to be sound, it has to be assumed that the three challenges have been adequately addressed. In turn, I showed that my analysis of an experimental result also elucidates the notion of an experimental artifact. In turn,

[24] Thus, while I am sympathetic towards Machery's (2020) analysis of replication as resampling, it does not address the question I am interested in here, namely the deep uncertainty (in practice) as to when the components of two experiments are in fact relevantly similar.

[25] See also Irvine (2021) for a similar argument.

once a given experimental result has been identified as an artifact of a failure to meet the reactivity challenge, this can give rise to an exploration of the specific ways in which the phenomenon under investigation is sensitive to moderators.

Let me briefly turn to the question of what is the scope of my analysis. While it has been presented here as an account of experimentation in psychology, many aspects of the analysis are probably not specific to psychology. Indeed, one might say that the only aspect that is indeed specific to psychology is that one needs experimental subjects to understand the instructions of the experiment and to be willing to follow them. It is precisely this feature of experiments involving humans that other authors have highlighted as potentially giving rise to a specific kind of reactivity, i.e., to artifacts that can result from the experimental subjects' awareness of being studied. I have argued that my account of reactivity not only subsumes this kind of reactivity as a special case but also enables us to develop a more general account of the very notion of an experimental artifact.

In addition to that, however, I want to highlight another peculiarity of the psychological subject matter, which makes its reactivity particularly hard to contain. The has to do with the fact that the subject matter typically has to be addressed at the "whole organism" level, i.e., as a black box that potentially has any number of complex (and interacting) entities and processes "inside" it. What I have in mind here is the fact that experimental manipulations are often intended as manipulations of something that is not directly accessible. Likewise, the variables whose reactivity has to be suppressed are often also not directly accessible. The same is true for the measurement of the intended dependent variable. Add to that the problem that many of the relevant processes and entities are not well understood, and it should be clear that the manipulation-, measurement- and confounder challenges pose huge problems to practicing experimental psychologists.

Acknowledgements I would like to thank Julie Zahle, Caterina Marchionni, and Marion Godman, as well as members of the audience at the workshop, "Reactivity in the Research Process" (Bergen, February 2020) for an inspiring workshop and helpful discussions. More recent versions of this paper were presented at Hannover University (April 2020), the Max Planck Cognition Academy (January 2021), the Ghent/Brussels Work-In-Progress-Colloquium (March 2021), and the philosophy colloquium at the University of Warsaw (April 2021). Many thanks to participants of each of these events. I am also grateful to Duygu Uygun Tunç as well as the two extremely helpful referees for this journal for their thorough and constructive feedback. Finally, and as always, the work of Jim Bogen and Jim Woodward has been invaluable in shaping the analyses presented here (though, of course, any errors and misunderstandings are all mine).

Availability of data and material (data transparency) Not applicable.

Code availability (software application or custom code) Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest/competing interests Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, XCVII (3)*, 303–352.

Boyle, A. (2021). Replication, uncertainty and progress in comparative cognition. *Animal Behaviour and Cognition, 8*(2), 296–304.

Canali, S. (2020). Towards a contextual approach to data quality. *Data*, *5*(4), 90. https://doi.org/10.3390/data5040090.

Colaço, D. (2018). Rethinking the role of theory in exploratory experimentation. *Biology & Philosophy Biology and Philosophy, 33*, 38. https://doi.org/10.1007/s10539-018-9648-9

Collins, H. (1985). *Changing order: Replication and induction in scientific practice*. Sage Publications.

Collins, H. (2004). *Gravity's shadow: The search for gravitational waves*. The University of Chicago Press.

Craver, C., & Dan-Cohen, T. (in press). Experimental artifacts. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/715202

Crasnow, S. (2016). Bias in social science experiments. In: McIntyre, L & A. Rosenberg (eds), *The Routledge companion to the philosophy of social science* (pp. 191–201). Routledge, London.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass Publishers.

Danks, D. & Eberhard, F. (2009). Conceptual problems in statistics, testing and experimentation. In J. Symons & C. Paco (Eds.), *Routledge companion to the philosophy of psychology* (pp. 214–230). Routledge.

Darwin, C. (1872). *The expression of emotions in man and animals*. John Murray.

Feest, U. (2011). What exactly is stabilized when phenomena are stabilized? *Synthese, 182*(1), 57–71.

Feest, U. (2014). Phenomenal experiences, first-person methods, and the artificiality of experimental data. *Philosophy of Science, 81*, 927–939.

Feest, U. (2016). The experimenters' regress reconsidered: Tacit knowledge, skepticism, and the dynamics of knowledge generation. *Studies in History and Philosophy of Science, Part A, 58*, 34–45.

Feest, U. (2019). Why replication is overrated. *Philosophy of Science, 86*(5), 895–905.

Guala, F. (2000). Artefacts in experimental economics: Preference reversals and the Becker-Degroot-Marschak mechanism. *Economics & Philosophy, 16*, 47–75.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70*, 1195–1205.

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

Guttinger, S. (2019). A new account of replication in the experimental life sciences. *Philosophy of Science, 86*, 453–471.

Hayes, A. (2014). *Moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.

Hon, G. (1989). Towards a typology of experimental errors. *Studies in History and Philosophy of Science, 20*(4), 469–504.

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science.* https://doi.org/10.1177/1745691620970558

Jimenez-Buedo, M. (2011). Conceptual tools for assessing experiments: Some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology, 18*(3), 271–282. https://doi.org/10.1080/1350178X.2011.611027.

Jimenez-Buedo, M. (2021). Reactivity in social scientific experiments: What is it and how is it different (and worse) than a placebo effect? *European Journal for Philosophy of Science, 11*, 42 (2021). https://doi.org/10.1007/s13194-021-00350-z

Jimenez-Buedo, M., & Miller, L. (2011). Why a trade-off? The relationship between the external and internal validity of experiments. *THEORIA, 69*, 301–321.

Jimenez-Buedo, M., & Guala, F. (2016). Artificiality, reactivity, and demand effects in experimental economics. *Philosophy of the Social Sciences, 46*(1), 3–23.

Jones, S. (1992). Was there a Hawthorne effect? *American Journal of Sociology, 98*(3), 451–468.

Lavelle, J. S. (2020). When a crisis becomes an opportunity: The role of replications in making better theories. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/714812

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science, 82*, 810–821.

Leonelli, S. (2020). Scientific research and big data. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (summer 2020 edition). https://plato.stanford.edu/archives/sum2020/entries/science-big-data/. Accessed 14 Sept 2021.

Lewis, J. (2020). Experimental design. Ethics, integrity and the scientific method. In R. Iphofen (Ed.), *Handbook of research ethics and scientific integrity* (pp. 459–474). Switzerland.

Machery, E. (2020). What is a replication? *Philosophy of Science, 87*, 545–567.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. The University of Chicago Press.

Mook, D. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology, 114*(5), 657–664.

Norton, J. (2003). A material theory of induction. *Philosophy of Science, 70*, 647–670.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Rasmussen, N. (1993). Facts, artifacts, and mesosomes: Practicing epistemology with the electron microscope. *Studies in History and Philosophy of Science, 24*, 227–265.

Reiss, J. (2019). Against external validity. *Synthese, 196*, 3103–3121. https://doi.org/10.1007/s11229-018-1796-6

Potters, J. 2019. Stabilization of phenomenon and meaning. On the London & London episode as a historical case in philosophy of science. *European Journal for Philosophy of Science*. https://doi.org/10.1007/s13194-019-0247-7.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy* Compass. 2019;14:e12633. https://doi.org/10.1111/phc3.12633

Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review, 40*(4), 337–354.

Rubin, M. (2020). Repeated sampling from the same population? A critique of Neyman and Pearson's responses to fisher. *European Journal for Philosophy of Science, 10*, 42. https://doi.org/10.1007/s13194-020-00309-6

Schickore, J. (2005). Through thousands of errors we reach the truth' -- But how? On the epistemic roles of error in science. *Studies in History and Philosophy of Science, 36*, 539–556.

Schickore, J. (2019). The structure and function of experimental control in the life sciences. *Philosophy of Science, 86*, 203–218.

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science, 64*, 65–74.

Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Front. Psychol., 8*, 702. https://doi.org/10.3389/fpsyg.2017.00702

Strack, F., Martin, L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777.

Sullivan, J. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese, 167*, 511–539.

Sullivan, J. (2015). Experimentation in cognitive neuroscience and cognitive neurobiology. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics* (pp. 32–47). Springer.
Tulodziecki, D. (2013). Underdetermination, methodological practices, and realism. *Synthese, 190*, 3731–3750.
Tunç, D. U. & Tunç, M. N. (2020). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. https://psyarxiv.com/pdm7y/. Accessed 14 Sept 2021
Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., et al. (2016). Protocol registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917–928.
Weber, M. (2005). *Philosophy of experimental biology*. Cambridge University Press.
Wimsatt, W. (1981). Robustness, reliability, and overdetermination. In R. Brewer & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 123–162). Jossey-Bass.
Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science, 67*, 163–179.
Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences,* 1–37. https://doi.org/10.1017/S0140525X20001685