Empirical Precision and Theoretical Depth Across the Sciences

Andrin Spescha
ETH Zurich
KOF Swiss Economic Institute
spescha@kof.ethz.ch

January 27, 2026

Abstract

The sciences differ in the precision of their empirical tests. A central determinant of this precision are the applied auxiliary hypotheses, which can encompass everything from theories to apparatus to data analysis. If they are subject to large variation within and between studies, the obtained results will vary, too. This paper investigates how the ways to handle auxiliary hypotheses differ across the sciences. This covers, for example, the possibility to separate and test auxiliary hypotheses, to reveal them through intervention, or to construct the experimental setup to exclude false ones. The paper focuses on a comparison of physical work in the laboratory in the natural sciences to data work with computers in the social sciences. The interaction with physical experimental setups allows natural scientists to better test, manipulate, and neutralize auxiliary hypotheses. In contrast, the collecting, processing, and analysis of data in the social sciences faces severe difficulties in choosing the right auxiliary hypotheses. Too many of them seem equally true. Social scientists thus struggle with numerous researcher degrees of freedom in their studies. Consequently, the natural sciences can better narrow down false auxiliary hypotheses than the social sciences, which allows them to achieve more precise empirical results and in turn reach deeper levels of theoretical development.

Keywords: empirical results - auxiliary hypotheses - natural and social sciences - experiment - data analysis - researcher degrees of freedom

1. Introduction

The sciences differ in the breadth and depth of their theories. Several empirical studies show evidence in line with such a hierarchy of the sciences (Comte 1908). The natural sciences at the bottom of the hierarchy achieve higher levels of consensus than the social sciences at the top (e.g., Fanelli 2010, 2012, Fanelli and Glänzel 2013, Simonton 2004, Lamers et al. 2021, Chen et al. 2018, Evans et al. 2016), implying that the former exhibit more fully articulated scientific paradigms than the latter (Kuhn 1962). One central reason for this difference is the precision of empirical tests (Kuhn 1961). Sciences that can produce more precise experimental results are able to develop theories of greater breadth and depth.

Empirical precision has traditionally been discussed in light of the Duhem-Quine thesis (Duhem 1906, Quine 1951), which states that an experimental test of some hypothesis requires all kinds of auxiliary hypotheses. Falsification of the main hypothesis is thus difficult, as some of these auxiliary hypotheses might be false, and not the main hypothesis itself. Such auxiliary hypotheses can refer to experiment, theory, but also basic assumptions like logic. Philosophers of science like Popper (1959), Kuhn (1962), and Lakatos (1978) were mainly concerned with the impact of the Duhem-Quine thesis on the falsification of the theory under investigation. That is, whether a falsification refutes the theory or whether the possibility of modifying the theory renders this difficult. In contrast, the focus in this paper will be on only the set of auxiliary hypotheses used in experimental test. This is in line with the "new experimentalists" in the philosophy of science (e.g., Hacking 1983, Ackerman 1985, Galison 1987, 1997, Giere 1988, Franklin 1989, 1990, Mayo 1996). These authors emphasize that in science providing reliable observational evidence is essential and therefore focus on experiment in all its forms.

The central difficulty expressed by the Duhem-Quine thesis is knowing which, if any, of the numerous auxiliary hypotheses might be false. Duhem (1906) argued that this process does not follow clear methods or rules. The false auxiliary hypotheses cannot be pinned down by logical analysis. However, in scientific practice, there are ways that make it possible to address the Duhem-Quine thesis. Mayo (1996) argues that scientists combat the Duhem-Quine thesis by actively searching for errors, or false auxiliary hypotheses. Scientists can narrow down false auxiliary hypotheses. However, this is not possible to the same degree in all sciences.

This paper investigates how conditions, strategies, and findings to handle auxiliary hypotheses differ between the natural and the social sciences. The comparison brings to light respective strengths and weaknesses that otherwise remain more hidden. To describe the natural sciences, the paper relies more on works from the philosophy of science, particularly from the new experimentalist. In contrast, to describe the social sciences, the paper relies more on works from the newly emerged field of metascience, which is the scientific study of science itself. The paper argues that whereas the natural sciences generally rely on more auxiliary hypotheses than the social sciences, they are nonetheless better able to narrow down false ones. They can benefit from several distinct aspects laid out in detail throughout the paper. Consequently, the natural sciences can put forward more precise empirical results.

Note that the concern of the paper is not specifically with the Duhem-Quine thesis; that is, when we have experimental evidence that is contrary to a theory's prediction. Instead, the focus is more broadly on how a lack of control over auxiliary hypotheses manifests in variation in experimental results. The latter hinders any kind of evaluation of theory. Variation in auxiliary hypotheses stands in-between theory and observation. It can block effective communication between the two and stall further development of theory.

2. The hierarchy of the sciences

2.1 Consensus and progress

The large majority of scientists live and work within the normal science of their respective scientific paradigms (Kuhn 1962). Normal science is cumulative. Over time, scientific theories become more and more articulated and match to nature at an increasing number of points with increasing precision (Kuhn 1970a). The famous scientific revolutions, where our most important theories completely change their form, happen only infrequently. The central aspect of a scientific paradigm is the consensus between scientists on what constitutes the fundamentals. Scientists do not argue about basic issues but rather build their research on them. Together the scientists can then investigate some field in a much higher detail. Kuhn (1962) even argues that progress is possible only after normal science has emerged.

The rigor of normal science will over time isolate severe anomalies that cannot be ignored. Kuhn (1961) emphasizes here the special importance of quantitative anomalies. In fact, measurement shows its greatest strength in anomalies. Quantitative anomalies are much harder to ignore than qualitative ones. Ad-hoc modifications of theories are easier to come by than ad-hoc modifications of precise numerical estimates. Numbers are neutral arbiters. Quantitative anomalies provide a "razor-sharp instrument" for the evaluation of a theory. They demonstrate deviations from theory with a strength that qualitative anomalies cannot imitate and are very difficult to explain away. Scientists are seldom willing to compromise the numerical accuracy of their theories. Quantitative anomalies therefore require looking for new qualitative phenomena. They are the unambiguous signals of crisis and at the same time provide the materials for revolution (Kuhn 1970b).

The high precision of empirical results in the now mature natural sciences Kuhn speaks of has sometimes created such quantitative anomalies that ultimately led to the demise of entire paradigms. Empirical results in the social sciences have so far never been that precise; they vary too much within and between studies. There are no important quantitative anomalies that show stable numerical discrepancies between both theory and empirical results. And if there appear indeed some such quantitative anomalies, they will not remain irrespective of further empirical tests, as they will be again subject to wide variation. Because the inconsistencies between theory and empirical results are not precise and stable in the social sciences, less clear indications for new theories arise. Without a clearly defined set of quantitative anomalies that resist resolution, social scientists have fewer reasons to transition to new paradigms.

2.2 The natural and the social sciences

The idea that the sciences are ordered in the form of hierarchy dates back to the sociologist Auguste Comte (1908). The sciences at the bottom are simpler and more general, whereas the sciences at the top are more complex and more special. The more complex sciences at the top thereby depend upon the more simple sciences at the bottom. Comte proposed to group the sciences in six primary divisions: mathematics, astronomy, physics, chemistry, biology, and sociology. This hierarchy of the sciences has been extended and reformulated in different ways over the past one and a half century.

However, the sciences also differ in their theoretical depth and empirical precision; that is, how far the paradigms in a scientific field have developed. In some of the natural sciences, theories and empirical facts develop closely together to create more knowledge. This knowledge then

constitutes the fundament for new inquiries. Scientists can discover theories that match the facts and then build on them to go deeper and discover new theories that again match the facts and so on. Over time the knowledge covers more areas in greater detail; theories become deeper and the empirical facts more precise. The sciences differ widely in this extension and reach of their respective paradigms.

Physics, for example, is a very mature field that has seen large progress over the past centuries. Until about the 1980s, physicists got ever deeper in their search for the fundamental building blocks of the universe. They went from matter to molecules, to atoms, to electrons, protons, and neutrons all the way down to the particles of the standard model. One reason for this development has been the high precision in experimental test, which allows identifying clear-cut empirical facts that constrain theories. The very high standards for precision show in the required statistical significance of up to six sigmas in particle physics, for instance. Of course, the research frontier in physics faces problems with numerous uncertain results, too, and estimates can vary widely. However, over time, important differences are often resolved. In the history of physics there has been strong disagreement about certain theories or experiments. But many times, they were eventually settled and knowledge increased (Franklin 2018).

In fact, one striking aspect of physics is that over time scientists do show consensus for important experimental results (Franklin and Perovic 2023). Similarly, in for instance laboratory biology we see that scientist manage to establish empirical facts that everyone agrees about (Latour and Woolgar 1979). This ability to achieve consensus on experimental results seems to be characteristic to some natural sciences.[1]

The social sciences, in contrast, have not managed to establish such layers of theories that accurately match empirical facts. Knowledge of the economy, society, or psyche still operates close to the surface. Theories do not become deeper and thus more closely aligned with more precise empirical facts. The resolution of important questions in economics, for example, has been slow and uncertain (Solow 1982). In none of the social sciences do we have clear-cut answers to the big questions. Instead, too many alternative theories compete to explain the same phenomena.

Even (non-trivial) empirical facts are hard to establish in the social sciences. Empirical studies cannot build on each other in the same way as they do in some of the natural sciences. When a set of empirical studies present estimates of some effect, they do not together establish some stable empirical fact that new empirical studies can incorporate and use as a steppingstone. The estimates are much too imprecise and vary too greatly. They do not converge to some precise estimate and thus offer no solid fundament to investigate the effect in more detail. New studies rather start anew with alternative estimates.[2]

---

[1] The deep, empirically precise theories of some natural sciences are not secure either. They may still turn out to be false after enough time has passed. But then the theories are usually replaced with even deeper, even more empirically precise theories. These are the revolutionary changes occurring in science from time to time (Kuhn 1962). They replace the paradigm dominant in a scientific field with an even more accurate paradigm of greater scope.

[2] The estimates of different studies do also not just average out to some empirical fact, because they are seldom directly comparable to each other. All studies investigate all kinds of alternative versions of the actual main hypotheses with all kinds of alternative empirical approaches in all kinds of alternative contexts. We thus do not know whether their different results emerge either from their different empirical approaches or from their different hypotheses or contexts. In practice, if findings between studies differ,

Hence, the research landscape in the social sciences grows broader, but the resolution remains about the same. New studies stay on similar levels of generality and do not become much more detailed. They cannot refine some previously established estimates and thus also lack systematic connection to them. They can complement aspects left out in previous studies. Yet the estimates from all the studies do not hang together to give rise to some coherent whole for deeper research. Studies in the social sciences tend to add-up and do not progress vertically but rather horizontally.

Notably economics does reach deeper levels of theoretical sophistication, too, but this development is largely decoupled from empirical facts. Economic theory is not grounded to the same degree in the empirical world as for example the theories in physics are. Economic models are idealizations that give intuition, they do not map the actual economy. The ability to obtain deeper and more precise levels of knowledge in the form of a match between theory and empirics has been most characteristic of physics in the past. In contrast, modern particle physics for instance suffers from a decoupling of theory and the empirical world, too, as experimentation has become more and more difficult or even impossible (Hossenfelder 2018). Things become too small or too distant to measure with simple means. Facts get sparse and a manyfold of theories takes over. It has run into similar obstacles as the social sciences.

Because the social sciences are less bound to empirical facts, they generate a larger diversity of different ideas. They can seem more creative than the natural sciences, which are much more bound to what empirical inputs dictate. Those researchers that can create what other researchers like become successful. This in turn is not necessarily what corresponds to the facts. Some fields in the social sciences may become very popular, even though they show little connection to the empirical world. The main findings in several such fields in psychology have collapsed during the replication crisis (see, e.g., Open Science Collaboration 2015).

2.3 The empirical evidence

There are a number of studies showing empirical evidence for the existence of a hierarchy of the sciences. Most of these studies rank the sciences in their ability to achieve consensus and thus to accumulate knowledge. A mature science shares a "common background of established theories, facts, and methods", which can in turn serve as the basis for further research (Fanelli and Glänzel 2013). Importantly, consensus is a necessary but not a sufficient condition for the accumulation of knowledge. Scientists can reach consensus on false research findings, too (see, e.g., Nissen et al. 2016).

For example, Fanelli and Glänzel (2013) use bibliometric data to show evidence for a hierarchy of the sciences. Fanelli (2010, 2012) further shows that in in the social sciences, researchers have more room to achieve favorable empirical results than in the natural sciences. Simonton (2004), Smith et al. (2000), Best et al. (2001), and Ashar and Shapiro (1990) use various indicators such as use of graphs to create composite measures that find evidence for a hierarchy of the sciences. Finally, Lamers et al. (2021), Chen et al. (2018), and Evans et al. (2016) use

scientists compare the studies and isolate those factors that differ between them. However, to what degree these factors influence the findings is extremely difficult to assess. One does not see from the published study itself the importance of design, analysis, setting, or data gathering. Hence, scientists have a hard time knowing which factors are in effect responsible for the difference in all the findings. They usually choose some plausible aspects, like the country where the study took place. Since these may not have been the relevant factors after all, the understanding of the literature is often false.

text mining to show that disagreement, uncertainty, and consensus in different scientific fields follow a hierarchy of the sciences.
Note that these studies are social science studies, too. The ranking of and the distance between the sciences therefore vary from study to study and are far from stable. As in most social science instances, it would be hard to pursue more detailed empirical research on the basis of the evidence these studies present. Nonetheless, for the purpose of the present paper, evidence for the existence of a hierarchy of the sciences is sufficient. We do not need a precise ranking with many details; only that the sciences do show some hierarchy in terms of consensus and thus likely also in accumulation of knowledge.

3. Two categories: manual vs. data work

To illustrate the category of the natural sciences, the paper will focus on certain core aspects of laboratory experiments in physics, both small and large as well as old and new. Physics has long had the lead in experimental science, pioneering new techniques, methods, and settings (Franklin and Perovic 2023). To shed light on the category of the social sciences, quantitative observational studies using (large) datasets in economics, sociology, and political science will occupy the center stage. Nonetheless, the paper will discuss aspects of experiments in the social sciences, such as laboratory studies in psychology or field experiments in economics. Moreover, to broaden the category of the natural sciences, the paper will also consider certain aspects of laboratory experiment in biology. Biology is a large science, and fields like evolutionary biology are probably closer to the social sciences. However, the laboratory experiment remains the most characteristic element of modern biology (Weber 2004).
The paper focuses on the key differences in the ways to handle auxiliary hypotheses between manual work with physical objects in the laboratory, most characteristic of the natural sciences, and data work in front of the computer, most characteristic of the social sciences. Of course, data has become very prominent in the natural sciences, too. It is, however, always connected to at least some interactions with the physical and seldom based on just analyzing (given) datasets. Nonetheless, the more a science relies on data work, the more it will resemble the social sciences. Conversely, laboratory and especially field experiments have prominently entered all the social sciences. They are in their nature closer to the natural sciences again and can benefit from many important ways to handle auxiliary hypotheses.

4. The number of auxiliary hypotheses

The more auxiliary hypotheses an experiment involves, the more likely it becomes that at least some of them are false, which can in turn invalidate the experimental results. For example, compare the heroic tales of Galileo's experiments with the inclined plane to the experiments conducted in modern particle physics. Whereas the former rely on only very few auxiliary hypotheses, which are part of a very simple setup and visible by eye to everyone, the latter involve millions of auxiliary hypotheses, which are embedded in extremely complex setups and understandable only to very specialized experts.
Even at the beginning of the twentieth century, experiments in particle physics involved only a few scientists in a laboratory, with simple experimental setups and relatively cheap equipment. This is in stark contrast to today with for example the Large Hydron Collider in Geneva, where

experiments involve highly complex machinery and thousands of scientists. Galison (1987, p.263) argues that in particle physics the "material basis for experimental work has grown literally to monumental proportions". Of course, such large-scale experiments always show a high division of labor. They are organized like some large corporation. The actual work is still done in smaller teams, they are just connected to an overall whole. The large-scale experiments contain within them many separate, much smaller experiments.
We have observed a growth in the number of auxiliary hypotheses across all sciences over the last decades. Experiments have increased in size and complexity in almost every field. Using only few auxiliary hypotheses may allow for higher precision, but today such experiments cannot tackle our most important research questions anymore. Interesting experiments in all sciences build on an elaborate network of crucial auxiliary hypotheses. The deeper scientists want to go in their inquiries, the more auxiliary hypotheses they need. Nonetheless, large and complex experiments are not doomed to deliver biased results. To the contrary, the following chapters will show that some experimental sciences can handle even very large numbers of auxiliary hypotheses. Much more important is the question of how different scientific fields can handle them.

5. The ability to test

5.1 Piecemeal

A first strategy scientists use to handle auxiliary hypotheses is to proceed piecemeal (Popper 1963). When scientists add new auxiliary hypotheses to their experiment, they may test them, alone or in conjunction with others. This is possible when scientists can proceed step-by-step in building their experimental setup. Similarly, before the trial, scientists may check all auxiliary hypotheses, and after the trial, especially when it is a surprising result, they may check them again. Of course, it is impossible to test every auxiliary hypothesis. Each inquiry where we try to test auxiliary hypotheses can only go so far that doubt becomes unreasonable. We must give up testing of the testing at some point, since otherwise it becomes infinite (Popper 1959).
Auxiliary hypotheses in the natural sciences are better separable and testable because many of them are physical in nature. In physics, for example, the auxiliary hypotheses are to a large degree machines, apparatus, instruments, detectors, or physical tools. They mediate between the microworld and the world of knowledge (Galison 1997). In biology, laboratories consist of a combination of biological materials, measurement instruments, preparation tools, etc. (Weber 2004). Scientists can take these apart and, if necessary, test each piece to receive evidence for their adequacy (Knorr Cetina 1999, p.57). This is much less feasible in the work with given datasets in social science studies. Scientists cannot just separate and test individual auxiliary hypotheses.
In data work, knowledge of the truth or falsity of most auxiliary hypotheses in the setup of a study is impossible. The true statistical model remains invisible. For example, how can one know whether the statistical model includes all relevant explanatory variables? There is no solid background which choices of auxiliary hypotheses can be compared against. To learn from errors, we need a hard surface against which we can identify errors in the first place. We need to be able to know whether a particular auxiliary hypothesis is false or not. This is difficult in data work, where scientists can often only argue for something but seldom test it. For most auxiliary hypotheses, scientists have no other way than to assume them as true. Of course,

scientists can give reasons why some important assumptions should hold: exogenous shocks, placebo tests, balancing tests, etc. The context of the study can provide such indications, making some settings more credible than others. Scientists thus provide arguments why some estimate is not endogenous, although they will not know its actual extent; some simple correlational effect may in fact be less endogenous. However, for the large majority of auxiliary hypotheses, providing reasons why they should hold is not possible. Unfortunately, imposing these other auxiliary hypotheses nonetheless has a strong impact on the results (as shown, for example, in multi-analyst studies like Silberzahn et al. 2015, Breznau et al. 2022, or Huntington et al. 2025).

Natural scientists can also better test their experimental setup in a piecemeal fashion against the relevant theoretical foundations in their field. This includes many well-established theories of the apparatus. For example, scientists can create detector response models to compare how many events a detector sees with how many it should see (Knorr Cetina 1999). Theories in the natural sciences are much better representations of the empirical world and can thus better inform decisions about auxiliary hypotheses. They are also much more consistent with each other and together forbid numerous steps involving false auxiliary hypotheses. Social scientists cannot do this in their studies to the same degree, as variation between all the various possible theories to test against is way too high. They usually have no theoretical foundations against which they can compare their intermediate steps and thus have difficulties knowing whether they are on the right path in data collection, processing, and analysis.

Experiments in the social sciences suffer from a similar problem. Important auxiliary hypotheses are here task, conditions, and rules of the game. They correspond to the apparatus. Because these auxiliary hypotheses are abstracted, it becomes hard to evaluate whether they are true or not. They are always false in a sense. At most they can be adequate models. This makes it difficult to test auxiliary hypotheses. Against what background can such auxiliaries even be true? The real world is explicitly excluded. The experiments create effects that rely on conditions that are never satisfied in social or economic life. The mechanisms there may be very different. Experiments in the social sciences thus offer a wealth of possible ways to create effects, but whether they indeed occur this way in the world is often open.

Overall, in the work with data, most auxiliary hypotheses remain glued together, and scientists must often take this network of auxiliary hypotheses as a whole without the ability to separate and test its elements. They cannot, for example, test for the correctness of the coding of the variables relevant for the data analysis. In general, the more physical objects and the less data an investigation involves, the better separable and testable the auxiliary hypotheses are. Such tests also often appear in separate scientific studies.

5.2 Calibration

A second strategy is to test the apparatus of the experiment on some known test object. Should the apparatus be able to correctly detect the properties of the object, scientists can infer that the apparatus operates properly, which in turn validates other results from the apparatus. It would be a large coincidence otherwise for the apparatus to correctly detect the properties of the test object (Hacking 1983). This strategy is widely used across the sciences and usually called the calibration of the experimental apparatus (Franklin 1989). Scientists often also test their apparatus on other apparatus that vary in attributes, use different techniques, or are based on distinct theories. It would again be a large coincidence for apparatus varying along

such dimensions to produce identical results (Hacking 1983). The idea is to deploy one set of auxiliary hypotheses to test another, separate set of auxiliary hypotheses.
The use of calibration to verify auxiliary hypotheses is crucial in the natural sciences. Scientists can test their apparatus on known results from past studies; things about which they already know the relevant properties, the more varied the tests, the better. Mayo (1996) argues that, in data work, scientists can make use of statistics in a similar way as with a physical instrument. They can apply it as tool to identify effects. However, scientists cannot identify the reliability of their statistical tools in the same way scientists with physical tools can. In the social sciences, we cannot know whether we have true effects. There is no repertoire of known results from past studies on which we can test our statistical tools to for instance calibrate them. The only method social scientists have in this respect is Monte Carlo simulations. However, they are artificial environments where we can never know whether real environments indeed represent those simulations. The assumptions Monte Carlo simulations rely on to test statistical tools are usually very strong.

5.3 Sub experiments

A third strategy is dividing the experiment into a series of smaller sub experiments (Galison 1987). The different sub experiments would need to be consistent with each other. Systematic variation of the respective experimental conditions will unearth artifacts by causing discrepancies between them. To separate an empirical study into mutually consistent sub experiments is widely spread in the natural sciences. Latour and Woolgar (1979) describe how in laboratories in biology scientists undertake many such sub experiments. Together they serve as mutual controls. The demand to pass through all of them reduces the probability of an artefact. Large-scale experiments in physics also run separate sub experiments that contribute to the overall understanding (Galison 1987). Scientists design their sub experiments to test certain confounding factors. Several sub experiments simultaneously allow cross-checking the different approaches (Franklin 1989). If they agree, all the better. If not, the sub experiments must be reconsidered in detail, since some of the applied approaches might not work correctly. Their agreement or disagreement delivers important information about the measured quantities. Experimental physicists sometimes even conduct "sister experiments", where they set up experiments independently from each other to compare their respective results (Knorr Cetina 1999). Natural scientists in general often justify the ceteris paribus clause in their studies by relying on past experimental results that rule out the influence of confounding factors, similar to a series of sub experiments conducted by third-parties (Mayo 1996). This is an aspect foreign to the social sciences. Empirical studies do not rely on other published studies to justify the ceteris paribus clause. More generally, social scientists do not pursue sub experiments in the form of additional studies to rule out some possible confounding factors. To the contrary, a study is usually seen as interesting if it shows the existence of some alternative factor, not if it rules it out. However, social science studies do triangulate results with different sub experiments within the scope of the same study, especially in psychology, where scientists pursue several sub experiments that target the same underlying theoretical explanation. In contrast, this is more difficult in observational studies in economics, sociology, or political sciences. These are usually restricted to a single or at least very similar quasi or natural experiments and cannot rely on additional sub experiments.

5.4 The results themselves

A fourth strategy is testing whether the results align with well-corroborated theories. Alignment of the observations with such theories provide reasons to believe in the observations. Conversely, using apparatus that rests on a well-corroborated theory provides apriori confidence in the apparatus itself (Franklin 1990). The results themselves can also speak for the correctness of a result, especially if they form a consistent pattern. A fifth strategy is relying on pre-trial or after-trial research on important components of the main experiment. Scientists can estimate the influence of a critical auxiliary hypothesis in some separate experiment. This helps ruling out potential problems. In some sciences, entire subfields concern themselves with the study of such problems.

The fourth and the fifth strategies are present across both the social and natural sciences. Results can speak for themselves in every context. Consider, for instance, the construction of the first cyclotron: „There is no doubt that the agreement between Lawrence's theoretical calculation and instrumental behavior was central to their confidence that the four-inch cyclotron was working properly. However, the phenomenon itself must not be lost of sight. Livingston observed a sharp, recognizable, and repeatable change in the collector current as he varied the magnetic field strength. The effect is just too dramatic to be noise." (Baird 2004, p.53). A series of results that are consistent with each other is very well possible in the social sciences, too, such as effect sizes that increase monotonically in the theoretically expected direction. On the use of pre-trial or after-trial research in particle physics, Galison (1997, p.429) comments that: "We have seen many new methods of avoiding misreading arise with the growing scale of particle physics. These include the development of subfields for the study and control of distortions, the understanding of personal error, and the avoidance of spurious ascription of patterns". The social sciences do extensive research on such distortions, errors, and statistical noise, too. For example, the nascent field of metascience or specific fields in applied statistics or econometrics. However, this research is seldom targeted at specific problems that appear in one large experiment.

5.5 Robustness checks

A fifth, overall strategy is an explicit search for errors in the auxiliary hypotheses (Mayo 1996). For example, scientists can vary critical auxiliary hypotheses in the experiment and see what happens, amplify potential errors and observe how the patterns change, or introduce some standard to the experiment and see how the results deviate from it. Such discrepancies may be very informative. They can separate artifacts from genuine effects. When the results of the variation in auxiliary hypotheses are similar, we gain some assurance that they pose no problem. We likely have no artifact. When the results are different, we may be able to quantify the part that is due to an artifact and subtract it out, or at least have an estimate of its impact on the measured effect. If that impact is too small, one can discount it.

The main tool of the social sciences to mitigate the problem of false auxiliary hypotheses in their data work are robustness checks. They embody Mayo's (1996) search for errors. Social scientists use them extensively. They can identify whether particular auxiliary hypotheses exert a strong influence on the results. If the results hold up to variation in crucial auxiliary hypotheses, scientists can rule out their influence and it becomes less important whether they are true or false. The more auxiliary hypotheses scientists can rule out, the more confidence

they can have in their results, since it eliminates them as alternative explanations. However, many auxiliary hypotheses cannot be varied because they comprise the fundament of the entire study and are beyond the scientists' control. To check their robustness, scientists would need to run a new study.

6. Intervention

A further way to test applied apparatus and thus the network of auxiliary hypotheses is intervention (Hacking 1983). Scientists can manipulate the object under investigation and predict what should happen. If the detected results align with the prediction, this is evidence for a proper working of the apparatus. In the same way scientists can also learn whether some apparatus does not work properly. Over numerous trials, they will develop close familiarity with how the apparatus works, so they can know whether something is off. The systematic error of the apparatus will reveal itself. Scientists learn step by step how the apparatus functions, by trying it on all kinds of different objects, until they can confidently separate the real structures from artifacts of the apparatus (Hacking 1983).

However, given that the apparatus works properly, intervention can also serve a further purpose: scientists can improve their knowledge of the structures or mechanism of the object. They can explore whether some manipulation of the experimental setup changes the object under investigation in the expected direction, by for example varying specific conditions. Repeated trials will show how the objects tend to change and reveal potential artifacts. If, on the other hand, the results all support each other and are consistent, scientists can rule out an artifact. Each manipulation that leads to the same results about the object eliminates some alternative explanation (Hacking 1983, Woodward 1989, Galison 1987).

Hence, the two aims of intervention are testing the apparatus and learning about the experimental outcome. Intervention is possible in every science that works with physical objects. For example, in a biology experiment, scientists can use different fixing methods to different cells, vary the environmental conditions, or check for implausible behavior under specific circumstances, and thereby always observe results before and after (Franklin 1989). If the results of the intervention are in line with the predicted outcome, this provides evidence for either a proper working of the apparatus or gives insights into the experimental outcome itself, such as a significant effect on the investigated physical object.

For example, if microscopes were in general producing false images of specimen, scientists would have noted this by experimenting with them. The vision through microscopes and the ensuing manipulation of specimen would have led to inconsistencies (Hacking 1983). Similarly, when using a microscope, scientists can vary the setup with which they investigate an object. If some aspect of the object remains the same under varying conditions, they can have more confidence in it.

The key to successful intervention is the ability to alter the experimental setup and to do so fast. While such repeated trials are well possible in the laboratory, this not the case in observational studies in the social sciences. Observational evidence happens only once. The data is given and the scientists have no way to see how it changes through intervention. In contrast, laboratory experiments in the social sciences have the possibility to intervene to some extent. Scientists can test alternative choices for the research design: the operationalizations of the concepts, the instructions, and, crucially, the design of the treatments. An artifact would

be a result that holds only for a narrow range of these possible design choices, while successive trials could lead the way to some robust overall design. However, in comparison with the rapid trials that scientists can run on a physical object in for instance a laboratory experiment in biology, this remains a slow process.
Finally, the social sciences do have the possibility of conceptual replication. If scientists vary the data analysis and the results change in line with theory, they can profit from the same underlying idea behind intervention. However, the applicability of this conceptual replication depends strongly on what the dataset offers. It needs to include important alternative theoretical concepts.

7. Skill

Making an experiments work is a difficult task (Hacking 1983). To produce or create phenomena in a stable way requires a lot of skill. Education in the laboratory is therefore mainly learning the ability to know when an experiment works, and how to put it right if it does not. A course in the laboratory where in experiments all goes right the first time would teach little about experimentation, since learning is not as great as if many things would have failed repeatedly. Scientists need to be able to make a distinction between an experiment that works and one that does not.
The key to learning in the laboratory is the replication of known phenomena. Aspiring scientists must know what results they are supposed to obtain. This allows checking whether the auxiliary hypotheses in their experimental setups have been valid or not. Getting good at generating known results teaches a lot about how to work with research objects. Scientists get a feeling for how to do experiments. They can embark on alternative ways to produce a phenomenon. Scientists thereby learn how to interact with nature, like kids on a playground. They also learn how to debug everything that is unusual. Such information is usually not in the published papers, but very crucial to make the experiment work (Hacking 1983).
In the natural sciences, young scientists are trained on important and successful past experiments that produce well-established results. This way they can learn to develop the necessary knowledge of whether an experiment has worked or not. Over time they become skilled experts. Consider, for example, Jean Baptiste Perrin, who in his studies of Brownian motion repeated nearly all the past empirical tests from other physicists he based his work on, because repeating and getting good at reproducing anticipated results thought him much about his experimental objects, it gave a certain "feeling" for them (Mayo 1996).
Well-established, known results are something the social sciences do not really have. Scientists cannot learn well with replicating past experiments because they do not know whether those studies in fact produced true results or not, and therefore whether their applied approaches are in fact correct. When they do replicate some past landmark studies, they have few ways of knowing whether they have learned the right things. Of course, in actual practice younger social scientists do learn many things from past experiments, too. However, they may just keep repeating the same mistakes again and again, while at the same time getting more and more certain about them. The social sciences can sometimes create illusory expertise. Certain setups and patterns of data in a social science study are surely more convincing than other setups and patterns of data, and scientists can study how to recognize them. But they can never be quite certain whether they just learned how to create or interpret elaborate statistical bias or noise.

Consequently, if experimental scientists work in an environment that has a strong basis of well-established results, they will become true experts with time, like craftsmen. In contrast, if the environment is mostly consisting of work with complex datasets, the skill they learn is much more ambiguous, and a lot if it will only be convention that is not said to better approach true results.

8. The malleable and the given

8.1 Create clean data

An important aspect to achieve precise measurements is the ability to control auxiliary hypotheses. Scientists can build the experiment in a way that excludes important confounding factors that may otherwise have a systematic influence on the results (Galison 1987). The idea is to build the experiments in a way that allows isolating some effect.
Natural scientists can often build the physical setup of their experiments in a way that excludes certain confounding factors and allows identifying the effect. They can make their auxiliary hypotheses true. Natural scientists can also test whether potential disturbances have an influence on the results. If they do, they can change the experimental setup to exclude them. In physics, for example, scientists can explicitly introduce confounding factors like electrical, magnetic, thermal, acoustic, or seismic disturbances to the experimental setup and measure their effects. If they have an influence, scientists can then proceed to create a more isolated environment for the experiment. Natural scientists are not faced with a given situation that can only be so good. They are much less bound to what is given to them than social scientists. They can invent their way out of some impasse and actively create the experimental setup. An adequate setup then also implies that the applied set of auxiliary hypotheses is true.
Consequently, while natural scientists build experiments that create their data in a way that suits them, social scientists must often take their data as given. In observational studies, researchers choose an economic, historic, or social situation such that their experimental setup becomes appropriate. Social scientists cannot just replicate the setup and vary only some aspect and hold all others constant to find out which aspect has a decisive influence. In the natural sciences, if two theories differ, scientists can sometimes create physical environments that embody the exact situation where the consequences of the two theories differ. In observational studies in the social sciences, scientists must look for where such a situation may have happened naturally. They cannot alter the situation in the way they desire. Social scientists can only search for datasets that meet the design of their experiments. This data will generally be less clean than any data tailored for answering a specific question. It limits the possibilities of what to ask.
Hence, a central problem with observational studies in the social sciences, both well-identified and not, is their reliance on given datasets. Of course, for each research design, scientists have a certain flexibility in how they adapt it to the specific circumstances of the data. Research designs in alternative contexts differ in their exact specifications, for example. Social scientists can build a strategy to identify causal effects with their data, too. However, they cannot do it to the same degree. The building of physical apparatus has numerous more possibilities to exclude different confounding factors than the work with data. In the latter, scientists must be content with that what nature offers them.

In contrast, field and laboratory experiments in the social sciences can certainly also profit from such an active creation of the experimental setup, since much of their implementation takes place in the physical world, too. In general, across all the sciences, while observational studies happen more in the data world, experiments happen more in the physical word. And the more a study rests on the physical, the better the control of the auxiliary hypotheses and the cleaner, that is, less confounded, the data becomes. We therefore have two overlapping advantages the physical offers in all scientific experiments. First, scientists can create their experimental setup to produce correct results. Second, scientists can also test whether their setup has produced corrects results. They can make the results, and they can also test them.

8.2 The use of statistics

In an experiment, scientists intend to distinguish real effects from artifacts. If they had perfect control, they would not need statistics. The measured effects are exogenous by construction. If the only way a change in the outcome can happen is through the treatment, and no other possible factors can influence it, one does not need much data analysis.

For example, in the laboratory in biology, scientists do not use much statistics because they instead rely on "control experiments" (Weber 2004). Scientists vary experimental conditions to eliminate disturbing causal influences. These methods are qualitative but still informative about how the experimental treatment works. Experimental biologists can causally intervene in their experiments. Control experiments show that a result does not stem from some other process, such as contamination. They allow checking whether the scientists have made some errors in their experiments. The more varied the experimental conditions, the more possible errors scientists can rule out. Because scientists can exclude errors this way, they need not to cancel them out through using statistics.

The use of control experiments and the ensuing absence of complex statistics implies that in laboratory biology bias in the estimated effects is generally low but noise can still be high. Even though biologists document in detail all the things they have used to build their experiments, results can still differ markedly between different experiments, because the investigated samples are often (very) small. They may contain only a handful of mice showing quite particular traits, for example. The produced figures, graphs, or tables can thus vary substantially between different samples. This is one important reason why many results in biology have not replicated (see, e.g., Errington et al. 2021).

Mayo (1996) argues that demonstrating the absence of errors in an experiment provides a severe test for the hypothesis under investigation. How to avoid errors, that is, how not to fall under the spell of the Duhem-Quine problem, is mostly discussed from the perspective of using the tools of error statistics. Mayo constructs error statistics as a broad account that consists of methods and models from classical and Neyman-Pearson statistics. She lays less emphasis on how scientists assure with the elegant physical design of experiments that they do not commit errors. Yet here lies a crucial difference between the natural and the social sciences. They both make use of error statistics, but only the former can devise severe tests. The latter suffer from all kinds of errors in their empirical tests. Good experimental design, strengthened by the use of reliable apparatus, assures high quality data, which in turn makes the use of error statistics straightforward and valid.

8.3 Simulations

In the early 1900s scientists could in their experiments readily build around some arising error. The comparatively small size and low cost of the used apparatus made it possible to quickly rebuild everything such that the error could be eliminated. This includes redesigning the experiment, vary the setup, or build additional devices (Mayo 1996). Such reconstruction made the data exogenous again. Today, in contrast, the apparatus is in many experiments, especially in particle physics, so large and expensive that reconstruction is an option only to some extent. This means that the data must remain more endogenous. Scientists therefore resort to running simulations to be able to nonetheless isolate the targeted effect (Galison 1987). Instead of changing the apparatus, they simulate changes of the apparatus on the computer. It allows scientist to see through the computer what would happen if they had actually changed the apparatus. For example, physicists use computer simulations to calculate how many observations would have happened through unwanted systematic factors. They can simulate hypothetical worlds where problematic confounding factors were at play. To simulate the behavior of big machines on the computer compensates for the lost ability to physically manipulate them. "The computer simulation allows the experimentalist to see what would happen if a larger spark chamber were on the floor, if a shield were thicker, or if the multitone concrete walls were removed" (Galison 1987, p.265). The computer simulation can create situations that cannot even exist in nature and allows investigating alternative universes. In such simulations, scientists work similar to actual experiments. They can vary the inputs of the simulations and compare the respective outputs with each other. If they observe stability, they know they are on the right path. Simulations lie somewhere between physical experiment and data analysis. The intense use of simulations in some parts of the natural sciences has moved them away from the benefits of the physical more towards the issues prevalent in data work.

9. Repeated runs

9.1 Selection of observations

Noise without systematic bias manifests in smaller-scale experiments in physics as imprecision in the experimental run. The same experimental setup delivers different results for each run. The run itself may thereby be only a single datapoint, in the form of a measurement but also a graph or a counter. Consider here, for example, the famous oil-drop experiment of Millikan, which measured the negative charge of a single electron (Ackermann 1985). Scientists can usually stop the experiment early or late or exclude some datapoints. They can thereby select favorably. How to decide between true and false observations? How to recognize erroneous experimental runs? When measurement is imprecise and the sample is small, scientists can simply choose those noisy versions of the experimental runs that look best for their purpose. This is a particular way of fitting noise.

A closely related issue in all experiments is the problem of the stopping rule (see, e.g., Franklin 1990). The obtained results may influence scientists to stop the experiment and give up looking for false auxiliaries. For example, they may stop the experiment as soon as it agrees with theoretical predictions. This is a problem, as there might still be more false auxiliaries. The experiment should be stopped when they are all correct, or as correct as possible, and not based on the obtained results. If an experiment is run only once and it exactly confirms a theoretical

prediction, it may still be that some of the auxiliaries are false and the result is an artifact. However, this requires quite some coincidence. The more auxiliaries are involved and the more precise the prediction, the larger the coincidence. If, on the other hand, scientists can adjust and play around with auxiliaries until the result fits, the smaller the coincidence and the likelihood of an artifact increases. The more auxiliaries are involved and the less precise the prediction, the higher the likelihood of some artifact.
Selection of observations is a large problem with small samples. In large datasets, however, it becomes less decisive, given that the variables are not fat-tailed (which is a distinct, but large problem in all the social sciences). Here the problem of the stopping rule is more important. In fact, a more general instance of the stopping rule has become very prominent as a central shortfall of data work over the last decades. We will discuss this in the next chapter.

9.2 Hacking p-values and research designs

In experiments, repeated runs are an advantage, as they can help identifying false auxiliary hypotheses. The question is thus not just whether a science relies on data or not, but whether the data responds directly to the runs. Manipulations of the experimental setup can produce a stream of always different data. In contrast, datasets in observational studies remain the same for all runs. They are not newly generated with the respective runs. This has the consequence that repeated runs do not benefit the quality of evidence but rather devalue it. This is the problem of p-hacking, or result-hacking more generally.
In data collection, processing, and analysis, scientists often use the researcher degrees of freedom available to them to search for some specifications that show their desired results (Simmons et al. 2011). This could be a theoretically predicted or more credible estimate but also a specification in which placebo tests hold up. If in such a search scientists target statistical significance, we speak of p-hacking. Scientists try to lower the p-values or shrink the confidence intervals in order to make their empirical evidence seem stronger. Since researcher degrees of freedom in data analysis but also in collecting and processing the raw data are always numerous, scientists have a lot of room to hack their results in a specific direction. Because the term p-hacking is more widely spread, we will use it instead of result-hacking.
We can further differentiate between fragile and robust p-hacking. Fragile p-hacking means that scientists search for some statistically significant specifications that would collapse if they changed only some minor researcher degrees of freedoms. In contrast, robust p-hacking means that scientists search for some statistically significant specifications that reside within an entire set of specifications that are all statistically significant. The results would also hold up if the scientists changed some more major researcher degrees of freedom. Robust p-hacking is a trial and error process where the theoretical arguments develop together with the empirical findings. It ends only after the scientists have built apparently elegant matches between theory and empirics. While fragile p-hacking is widely regarded as problematic, robust p-hacking is a much more accepted practice.
Any type of p-hacking faces two distinct problems (Spescha 2021). First, searching through researcher degrees of freedom may fit statistical noise and this way create an artifact. Repeated runs invalidate the p-value, the filter that separates signal from noise. This is called the multiple comparisons problem. With statistically independent runs, for example, the actual p-value would about half each time. Second, researcher degrees of freedom are not all equally valid. P-hacking can increase the problem of false choices of researcher degrees of freedom because

statistical significance has the lead and not the theoretical implications. Hence, scientists mold their statistical model into a mixture of the true effect size together with noise and bias.[3]
A related problem is research design hacking, where scientists misuse the fact that the experimental setups often have many alternative implementations and none of them are clearly superior. The researcher degrees of freedom happen not in the data analysis stage but in the setup of the experiment. Some versions of the research design might product an effect, others not. Scientists can then present those research designs that produce favorable results. Of course, if the effects move in line with theoretically meaningful variation of the research design, it will not be a problem but actually an improvement. In contrast, if the effects move in line with seemingly irrelevant aspects of the research design, the produced evidence will be weak. Note that in observational studies, the research design is given and less hacking takes place. However, the uncertainty behind it is equally large or even larger. One just does not observe it.
A complete theory would for every context describe the experimental setup or research design and the required data analytic decisions. Unfortunately, we never have such a theory. Scientists can thus exploit variations in research design, data analysis, and context to their advantage. They can hack all three: the experimental setup in the laboratory, the data analysis in observational studies, or the context in field experiments.

9.3 Some remedies

P-hacking is widely spread in the social sciences (see., e.g., Brodeur et al. 2016, John et al. 2012, Necker 2014, Banks et al. 2016). Unfortunately, while replication using alternative specifications can uncover fragile p-hacking, this is less feasible with robust p-hacking. P-hacked results are robust by their very design. They have been chosen because they hold up to other specifications. The results of some robustly p-hacked specification change only little if one varies a single researcher degree of freedom. Of course, in many cases already modest variation of only two or three researcher degrees of freedom together can still falsify such results. Nonetheless, to uncover robustly p-hacked results, new data is necessary. This is not possible in observational studies, and cost and time intensive in experiments. We thus do not know to what degree the social sciences are plagued by robust p-hacking. However, we have indirect evidence, because meta-analyses show that in most studies statistical power is too low and effect sizes are overestimated (see, e.g., Ioannidis et al. 2017). Robust p-hacking is a way to nonetheless find statistical significance in samples that would otherwise show nothing in particular. Because it capitalizes on statistical noise, effects are generally overestimated.
P-hacking happens in the natural sciences, too. It is a central reason for the replication crisis in both medicine and biology. In many of those fields, samples are small and noise is high. For example, scientists tend to take those images where results are strongest and best fit their theory. This selection on favorable outcomes is similar to the overestimation of effect sizes in p-hacking.
In the large-scale experiments of particle physics, data analysis has become very important and exceedingly complex. It makes the data analysis in the social sciences seem simplistic. However,

[3] In data analysis, scientists usually give room to the effect they want to identify. They unearth a robust relationship for that one effect that takes center stage. Had they had another focus, and the effect would only have been a side-story, it would have been much less consistent and less pronounced over all results. The scientists diminish other effects in the data on behalf of the one they want to identify. For example, if you read the online appendix of a study instead of the main text, results will generally be much weaker.

experimental physicists combat the arising problems more actively (Franklin 2018). Data analysis in many high energy physics experiments uses „blind“ techniques, where they agree on the steps of the data analysis before looking at the data. Alternatively, they divide themselves up in separate teams, each using their ideal methods on the data, and then reveal all results jointly. In the last decade, the social sciences have seen many initiatives going into this direction, too, with pre-analysis-plans or registered reports. However, they are still far away from being the standard.

One solution to p-hacking and research design hacking would be to test and then present as many specifications and research designs as possible. This would provide an overview of alternative scenarios. Whereas showing many research designs is possible only in some small-scale experiments, showing alternative specifications is always easily feasible. Presentation of numerous different specifications in some aggregated way would help the credibility of most studies substantially (see, e.g., the specification curves of Simonsohn et al. 2020).

10. Constructing the apparatus

Constructing physical apparatus is a complex craft requiring substantial effort. To obtain reliable results, much repetition is required. Scientists usually select some early line of experimentation that seems promising. They then constantly improve the experimental setup in a trial and error way (Ackermann 1985). Insights that occur during experimentation are incorporated. Scientists discover false auxiliary hypotheses and replace them with true ones. Overall this leads to a convergence toward an apparatus that gradually produces cleaner and thus better sets of data. The measurement becomes more and more precise.

The first prototype of a new apparatus is always crude. Arthur (2009, p.133), for instance, describes how Lawrence's first cyclotron used "a kitchen chair, a clothes tree, window glass, sealing wax, and brass fittings." It is already a success if the prototype works at all. The key is that scientists can successively improve it. Collins (1992) gives the example of a scientist building a prototype of a TEA laser, where it first showed many anomalies that only trial and error could fix. The physical theories that explain the function of most prototypes are often rather basic. Experimental apparatus is thus seldom derived from some physical theory but instead the product of continuous improvement.

Giere (1988) argues that the Duhem-Quine problem can be resolved by technology. Background knowledge in scientific testing does not consist exclusively in theoretical terms as hypotheses previously confirmed, that is, propositional knowledge. Some background knowledge is better thought of as embodied knowledge, in the technology used in performing experiments. The physical existence of the technology in the laboratory is thus a quite different dimension of scientific progress. It corresponds to knowledge embodied physically in the various research apparatus, which provides it with an exceptional reliability. Baird and Faust (1990) even describe scientific progress in the form of the accumulation of new scientific instruments. Today we have vastly greater ability to manipulate, control, and measure nature by means of vastly increased arsenal of scientific instruments.

11. Improving of the experiment

11.1 Narrow down false auxiliary hypotheses

We have seen that the open-endedness of the physical world offers more room, or options, to generate setups that produce unbiased effects than the data world. Scientists can build around arising errors. However, the physical world may offer so much room that it becomes a disadvantage. The open-endedness of the physical world is often orders of magnitudes larger than the openness of the data world. Physical setups incur thus also plenty more degrees of freedom than data setups. The question is thereby how to arrange the auxiliary hypotheses in the wake of the design of the experiment; that is, how exactly the parts of the experimental apparatus are built together. Scientists could in principle construct any type of highly original but also totally false experimental setup.

Nonetheless, in physical setups, there is usually a rather tight link between theories, apparatus, and research objects. Setups in the data world remain much more indeterminate. Statistical theories, theories of the object, and statistical modelling can map in numerous possible ways. The difference is that the more physical world of the natural sciences allows better narrowing down the false auxiliary hypotheses.

In physical setups, scientists can make use of all the discussed ways to rule out false auxiliary hypotheses: a) separate the parts of the apparatus and test them, using for example calibration, b) intervene by varying the experimental setup to uncover errors, c) learn experimental skill by replicating well-established findings, d) build around sources of error learned discovered during experimenting or from previous experiments, e) repeat the experiment to better separate signal from noise, and f) construct tailor-made apparatus. Scientists can do this until they are confident that they have eliminated all false auxiliary hypotheses. Together the different ways represent a potent arsenal to correct false auxiliary hypotheses.

In setups with data, in contrast, scientists have fewer of these means to identify and correct false auxiliary hypotheses. The applied methods have much less bite. Social scientists can certainly refine their empirical strategies and try to better isolate effects. But data setups in the social sciences suffer from a fundamental inability to verify most choices of auxiliary hypotheses. They can go over them repeatedly, but they have few means to test whether any choice is superior to another. Their refinements could have worsened instead of improved the study.

11.2 Too much freedom

Because false auxiliary hypotheses are harder to narrow down in the work with data, the consequence are numerous researcher degrees of freedom in data collection, processing, and analysis. Depending on the respective choices they cause substantial variation in results. This crucial problem has become more and more evident over the last two decades, especially in the social sciences. In most empirical studies, scientists face a veritable “garden of forking paths” (Gelman and Loken 2013), where they have too much room in choosing between different ways to collect, process, and analyze their data.

Researcher degrees of freedom are auxiliary hypotheses that all seem equally true but are in many instances not. Some of the available researcher degrees of freedom might be more, some less adequate. They merely imply that scientists do not have sufficient information to verify which of them are the correct ones. Scientists have no ways to assess their choices of researcher

degrees of freedom. Which path they will ultimately take instead depends on their respective information, previous experiences, or beliefs. Different scientists will choose differently, leading to potentially large variation in results for the same main hypothesis. Many choices of researcher degrees of freedom are also conventions that have developed over time. Scientists can become more and more confident about choices that need not be correct. They may enforce certain errors in their data work repeatedly without being aware of it.

By now the literature contains many multiple analyst studies that illustrate the problem for different scientific fields: economics (Huntington-Klein et al. 2021, 2025), sociology (Breznau et al. 2022), finance (Menkveld et al. 2024), psychology (Silberzahn et al. 2018, Starns et al. 2019, Schweinsberg et al. 2021, Hoogeveen et al. 2023), psychiatry (Bastiaansen et al. 2020), ecology (Gould et al. 2023), and neuroscience (Botvinik-Nezer et al. 2020). These studies do not just show the presence of many researcher degrees of freedom in the work with data, but also that these give rise to sizeable differences in the obtained results. The more complex the study becomes, the more pressing the problem, as researcher degrees of freedom multiply very fast.

A similar problem arises for the setup of experimental studies in the social sciences. In this case, scientists can create the context of their studies and thus answer the theoretical hypothesis they want, if feasible. However, theoretical hypotheses in the social sciences are very imprecise. They do not map the empirical world very closely but are often abstract models. Each theoretical hypothesis thus allows for the creation of numerous alternative experimental designs. Hence, it again entails degrees of freedom. Scientists do not have the necessary information how theory and context should connect to tie down all the alternatives. Nonetheless, how the experiment and treatments are designed and concepts operationalized and measured can greatly influence results. The chosen experimental setup may not generalize to other, equally relevant setups and changes in setups can lead to large variation in results.

For example, experiments in psychology are often quite particular, specific implementations of some theoretical, usually verbal, claim (Yarkoni 2020). The experiments are implicitly thought to hold over many more alternative implementations of the general theoretical idea. If one adequately incorporated all those various alternative implementations, the uncertainty intervals of the estimates would be many times larger. Factors that produce variation in estimates are subjects, stimuli, task, instructions, site, experimenter, culture, and such seemingly insignificant factors like weather, etc. Collectively they contribute large variation. Yet most experiments in psychology proceed as if they would not vary.

In field experiments in the social sciences, the setup to produce the effects is less abstract than in laboratory experiments. The scientists look at real environments with real treatments. They may be smaller and cover only a subset of all environments or treatments. Nonetheless, the treatments themselves are actual implementations in the worlds of politics, economics, or society. However, the correspondence between the respective treatments and theory remains loose. Many treatments from many different contexts could account for the same theoretical hypothesis.

Consequently, experimental designs in the social sciences, both in the field and in the lab, suffer from many researcher degrees of freedom, too; neither theory nor the actual context can pin them down sufficiently. Experiments push researcher degrees of freedom from the analysis stage toward the setup of the entire experiment. In fact, variation in experimental design can be tested by running the same hypotheses with different experimental setups on different random subsamples. Overall, the two sources show a similar impact on variation of results (Holzmeister

et al. 2024). As long as the social sciences lack the empirical methods necessary to align auxiliary hypotheses with the right theories or contexts, such variation will be an integral part of these disciplines. The large variation of estimates visible in the various multiple analyst studies is itself evidence that they have not yet found those methods.

12. Structure in the chain of auxiliaries

Hacking (1988, 1992) argues that the laboratory sciences consist of the following 15 elements, divided into three parts: 1) *ideas*: questions, background knowledge, systematic theory, topical hypotheses, modelling of the apparatus. 2) *things*: target, source of modification, detectors, tools, data generators. 3) *marks*: data, data assessment, data reduction, data analysis, interpretation. Variation in estimated effects can come from all these 15 sources. All of them are plastic, that is, scientists can alter them.

In observational studies, we have only *ideas* and *marks*. This could be seen as an advantage, that is, less auxiliary hypotheses that could be false and thus less worries about potential variation. However, the *things* bring a structure to the experiment that simplifies both *ideas* and *marks*. They can reduce theoretical problems (e.g., exclude confounding factors or other sources of error) and statistical problems (the data are tailored to the research questions at hand). The *things* bring both *ideas* and *marks* together more directly.

13. Well-established scientific theories

Together the different ways to handle auxiliary hypotheses can deliver a state of unproblematic background knowledge, such that scientists can trust the results of their studies. Importantly, once the new results become themselves well-established, scientists may use them in their investigations in other new studies. The results become themselves auxiliary hypotheses and part of the unproblematic background knowledge. Over time, this process may repeat in a virtuous cycle. Altogether this creates scientific progress with results that go ever deeper.

13.1 Replication for a cumulative literature

Replicability of experimental results is a must. Without it, experimental results cannot serve their role as the guiding instances in empirical science. For example, Popper (1959) argues that only replicable effects can falsify or corroborate a theory. Widely varying experimental results cannot do so. He argues that we should only take observations as scientific if we have tested them repeatedly. Only replicability can convince us that we are not dealing with a mere coincidence. Otherwise, science cannot get off the ground.

Unfortunately, all scientific fields contain non-replicable or invalid results. However, some do so more than others. Psychology and medicine have been hit particularly hard by the replication crisis. The main reason here is that one can actually replicate their experiments and find out whether they uphold or not. These fields often rely on experiments that are comparatively easy, cheap, and fast to replicate. In contrast, the other social sciences did not have a substantial replication crisis because most studies cannot even be replicated. They are snapshots of unique places and times, not replicable experiments. The most one can do is reanalysis. To replicate important quasi or natural experiments, scientists must rely on other contexts. Hence, in much

of the social sciences, we do not even know the extent to which studies are non-replicable or invalid. One might suspect that the state is at least as bad as in psychology and medicine.
Because only few studies are replicated, it can happen that some false study nonetheless gathers hundreds of citations. Due to the abundant researcher degrees of freedom, it may even gather many conceptual replications. In contrast, the cumulative nature of many natural sciences requires that important studies are replicable (see, e.g., Peterson 2025). In laboratory work in biology, when scientists publish something important, other scientists replicate it since they want to use it for their own work. Scientists first need to know whether the relevant findings are reliable. They thus conduct replication of others' works in their own laboratory, to see whether they can indeed build their own research on them. This means that popular methods or findings are regularly checked because they are used by others. If a study does not replicate, the community will get to know this. The possibility to use the results of others is a strong incentive to redo important studies. The key is therefore the cumulative nature of laboratory biology, how it builds on previous work and then goes deeper. Some studies establish a new fact, which raises a new question. This question is in turn answered by a new fact, which again raises a new question and so on. Studies build on the results of previous studies. Even within the same project in a laboratory, biologists use control experiments to verify the facts they think they know, often together with new assumptions. If some of those control experiments give different results, the scientists will have to go through everything again to track down the problem. They redo and retest all things. Knowledge is therefore cumulative within as well as across scientists.
The social sciences do not require that a study is replicable. This mechanism of control is mostly absent. It does not matter for the estimate of a treatment effect in your context whether others have estimated different treatment effects in their work. Empirical estimates in the social sciences do not directly build on each other. In the social sciences, due to the countless researcher degrees of freedom, any replication will anyway turn out differently and scientists do not really want to open this box of pandora. The absence of much replication makes it necessary to rely on only the published empirical findings. These are incorporated into subsequent research. If the estimates are generally precise, subsequent research will not be lead too far astray. In contrast, if they vary highly, later research will vary, too.

13.2 Bandwagon effects and convergence

Franklin (2018) argues that successful replication is a goal of experimental physics. Important results are likely to be replicated, while less important results may often not be so. Physics has a tradition of rigorous replication, especially for groundbreaking discoveries. Because it has such a strong theoretical fundament, when empirical studies show new discoveries that go against this fundament, it raises skepticism among scientists and the studies get extra scrutiny in replication attempts.
Of course, in the measurement of physical quantities, such as physical constants and properties of elementary particles, estimates can vary in size, too (Franklin 2018). Sometimes large changes occur, which may be due to smaller systematic errors, corrections of older experiments, or smaller errors in general. Sometimes flukes can happen, too. In fact, experimental results often show bandwagon effects, where they tend to agree more with previous measurements or with theoretical calculations, and only over time converge to some specific value measured with greater precision. One reason for this pattern is violation of the stopping rule (Franklin 1989).

Estimates move toward the true value, but in small steps, as experimenters do not want to deviate too strongly from the previously measured results.
Bandwagon effects are quite common in particle physics (Franklin 2018). The crucial point is that, over time, the measurements do (slowly) converge toward the true value. The results of studies in the social sciences do not show such bandwagon effects. Estimates from empirical studies do not converge toward some specific value. Instead, they vary quite unsystematically and scientists do seldom come to an agreement about the precise value of some effect.[4]

13.3 The construction of facts

If there has been much research in the past, and the research led to well-established scientific theories, scientists can use them for their experiments and take them for granted. They count as unproblematic background knowledge. Giere (1988) gives the example of protons as research tools. They are part of the technology that investigates nuclear structure, for instance. Scientists use the proton in so many alternative ways during their experiments that the theories describing them become almost trivially true for them. The same holds for the investigation of many research objects, like some biological organism. Past research has established many parts or mechanisms of the object, which can then serve as at least approximately true auxiliary hypotheses for further investigations. For example, genes in cell biology have such a function. They have become useful tools. Scientists use genes to learn about the processes under investigation, by for instance using them to manipulate outcomes. The genes have become knowledge that is so well-established that they become an instrument themselves.
Latour and Woolgar (1979) describe how scientists in a laboratory in biology construct scientific facts. They differentiate between five levels of facticity in scientific papers. The highest level are statements so persuasive that no reference is needed. They are recognized as facts by everyone in the field. Middle-level facts are statements that need citations to other scientific papers. The lowest level are statements in the form of conjectures or speculations. They appear most common in the conclusion. Scientific activity itself causes some statements to move up and others down this ladder. It does so by supporting certain statements with figures, diagrams, and statistics. Among the large numbers of statements scientists produce, a mere fraction becomes a fact, whereas all other statements stagnate, as no scientists take them up. Once a fact is established, scientists no longer contest it. Instead, it becomes the basis of further discussion and disappears from daily scientific activity and enters textbooks.
There is no such ability to construct empirical facts in the social sciences. At most, scientists can try to weigh empirical evidence for some effect from different studies against each other. They can then form a judgement about what the effect might be. But it does not become an empirical fact that is recognized by everyone in the field. The findings that enter textbooks are theories, often even based on mathematical proofs, but not empirical facts. The former are

[4] Replications of experiments in physics are often not exact, that is, identical attempts. Such exact replications happens mostly when a result seems hard to believe to the community. Instead, scientists try to do the same thing better, such as to produce less noisy or more stable effects (Woodward 1989). Scientists thus use different, and if possible better apparatus. This broader type of replication happens in the social sciences, too, where scientists try address the some effect with better methods. However, studies with only improved method are quite rare. Hypotheses and contexts are usually different, too. The studies are therefore generally too distinct to count as replications. Most often such studies are more like series of alternative measurements, and none are clearly superior to the others.

elaborate hypothetical constructions, and not empirical facts like in cell biology, for instance, about what a cell is made of, how it functions, and what it does. In biology, years of controversy between scientists sometimes crystallize into the construction of some unambiguous and uncontested empirical fact. Most of the time such controversy leads nowhere, but sometimes it does. Unfortunately, in the social sciences, it almost never leads to such an outcome.

14. Conclusion

We have seen that the sciences differ substantially over several dimensions in their ways to handle auxiliary hypotheses. Each scientific field would score differently on each of them. Consequently, to construct an exact ranking is difficult. However, we can point out some main lessons from the preceding discussion.

While we can find studies with any number of auxiliary hypotheses in any field, the natural sciences tend in general to rely on more auxiliary hypotheses. Their experimental setups are usually larger and more complex than the data collection, processing, and analysis in the social sciences. Nonetheless, the natural sciences produce more precise experimental results. Hence, the question opens us why this is so.

First, the auxiliary hypotheses in the natural sciences are mostly physical in nature. Scientists can take them apart and, if necessary, test each piece to receive evidence for their adequacy. For example, they can calibrate their apparatus on known results or cross-check it with other instruments. In contrast, in the social sciences, auxiliary hypotheses cannot be tested against some solid background. Instead, scientists must simply assume most auxiliary hypotheses. They can give evidence only for some more important auxiliaries.

The main tool in the data work of the social sciences are robustness checks, where scientists vary auxiliary hypotheses to observe their influence on the results. They are heavily used in all studies. However, robustness checks can only show that some auxiliary hypothesis has little influence on the results, but they can do nothing about it if it has large influence.

Second, intervention is much better feasible and much faster in the laboratory. Scientists can vary their experimental setups to exclude artifacts of the apparatus, for instance. Such fast intervention altering the entire setup is not possible in the social sciences, where scientists are usually bound to the respective quasi or natural experiment. It is possible in experiments in the social sciences, yet still orders of magnitudes slower than in the laboratory.

Third, repeating experiments from past studies builds up important skills in the natural sciences. However, this requires a repertoire of well-established past results. This is less feasible in the social sciences, because well-established results are scarce. Scientists therefore do not know whether they may in fact have learned some false approaches.

Fourth, whereas natural scientists can build experimental setups that exclude false auxiliary hypotheses, observational studies in the social sciences do not have much room for altering their research designs. They can address minor shortcomings, but the environmental context defines whether the research design is appropriate or not. In contrast, this does not hold for experiments in the social sciences. They can certainly build around sources of error, too.

Fifth, in experiments in the natural sciences, the data reacts to alternative runs of the experiment. Variation in the setup also changes the produced data. Especially smaller experiments can be repeated until the data is satisfactory. In contrast, observational studies, but also most field experiments, cannot be run more than once. Scientists usually run repeated

data analysis though. However, this does not improve the data but rather devalues it. Scientists can use the researcher degrees of freedom available to them to show favorable results. They may mold their empirical specifications into a mixture of statistical noise and bias.

Together these (and more) aspects allow the natural sciences to better narrow down false auxiliary hypotheses. In contrast, in the social sciences we observe potentially large variation in auxiliary hypotheses, both within and between studies, leading to large variation in the observed empirical results. The estimated effects do not converge to some value over time, like in some of the natural sciences. One key difference between the natural and the social sciences is therefore that the former can sometimes establish new empirical facts. These in turn serve themselves as auxiliary hypotheses in even deeper inquiries that again establish new empirical facts. An overall virtuous cycle takes place. Consequently, the natural sciences are much more cumulative than the social sciences.

Data analysis has become more important across all sciences. In modern particle physics, for example, apparatus has become too large, complex, and expensive to systematically vary it. With less control over the physical apparatus, control over data has become more and more central, in sometimes very complex ways. The experiments have in some instances lost contact with the physical world entirely and data analysis has often become the experiment itself (Galison 1997). This shift from the physical work to data work has introduced an extra dimension of variation that has been absent in the experiments of the past. In contrast, in the social sciences, data analysis has replaced less precise methods and has rather reduced variation in results. It offers a more condensed window to the social world than the qualitative work that dominated earlier periods. In any case, with data analysis in all its various forms becoming more and more important, the sciences will move closer together, from the physical and the qualitative towards more data work. Similar problems will thus show up in all of them. Data analysis adds a layer of variation to experimental results in all sciences. In fact, scientific progress seems to have slowed down across all fields. Studies have become less disruptive over time (Park et al. 2023). One reason for this development may be that we cannot go physical enough anymore and must rely too much on data analysis. It causes an imprecision in empirical investigation that makes it more difficult going broader and deeper with our theories.

15. Bibliography

Ackermann, R. J. (1985). Data, instruments, and theory: A dialectical approach to understanding science. Princeton University Press.

Arthur, W. B. (2009). The nature of technology: What it is and how it evolves. Allen Lane.

Ashar, H., & Shapiro, J. Z. (1990). Are retrenchment decisions rational? The role of information in times of budgetary stress. The Journal of Higher Education, 61(2), 121-141.

Baird, D. (2004). Thing knowledge: A philosophy of scientific instruments. University of California Press.

Baird, D., & Faust, T. (1990). Scientific instruments, scientific progress and the cyclotron. The British Journal for the Philosophy of Science, 41(2), 147-175.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. Journal of Business and Psychology, 31, 323-338.

Bastiaansen, J. A. et al. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. Journal of psychosomatic research, 137, 110211.

Best, L. A., Smith, L. D., & Stubbs, D. A. (2001). Graph use in psychology and other sciences. Behavioural processes, 54(1-3), 155-165.

Botvinik-Nezer, R. et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. Nature, 582(7810), 84-88.

Breznau, N. et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Proceedings of the National Academy of Sciences, 119(44), e2203150119.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. American Economic Journal: Applied Economics, 8(1), 1-32.

Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. Journal of Informetrics, 12(1), 158-180.

Collins, H. (1992). Changing order: Replication and induction in scientific practice. University of Chicago Press.

Comte, A. (1908). A general view of positivism. Tanslated by J.H. Bridges. Routledge.

Duhem, P. (1906). Physical theory and experiment. In Can Theories be Refuted? Essays on the Duhem-Quine Thesis (pp. 1-40). Dordrecht: Springer Netherlands.

Evans, E. D., Gomez, C. J., & McFarland, D. A. (2016). Measuring paradigmaticness of disciplines using text. Sociological Science, 3, 757-778.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. Elife, 10, e71601.

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. PloS one, 5(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics, 90(3), 891-904.

Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. PLoS one, 8(6), e66938.

Franklin, A. (1989). The neglect of experiment. Cambridge University Press.

Franklin, A. (1990). Experiment, right or wrong. Cambridge University Press.

Franklin, A. (2018). Is it the ‘same’ result: Replication in physics. Morgan & Claypool Publishers.

Franklin, A., & Perovic, S. (2023). Experiment in physics. The Stanford Encyclopedia of Philosophy.

Galison, P. (1987). How experiments end. University of Chicago Press.

Galison, P. (1997). Image and logic: A material culture of microphysics. University of Chicago Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, 348(1-17), 3.

Giere, R. N. (1988). Explaining science: A cognitive approach. University of Chicago Press.

Gould, E., et al. (2023). Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.

Hacking, I. (1983). Representing and intervening: Introductory topics in the philosophy of natural science. Cambridge University Press.

Hacking, I. (1988). On the stability of the laboratory sciences. The Journal of Philosophy, 85(10), 507-514.

Hacking, I. (1992). The self-vindication of the laboratory sciences. Science as practice and culture, 30.

Hoogeveen, S. et al. (2023). A many-analysts approach to the relation between religiosity and well-being. Religion, Brain & Behavior, 13(3), 237-283.

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2024). Heterogeneity in effect size estimates. Proceedings of the National Academy of Sciences, 121(32), e2403490121.

Hossenfelder, S. (2018). Lost in math: How beauty leads physics astray. Hachette UK.

Huntington-Klein, N., et al. (2021). The influence of hidden researcher decisions in applied microeconomics. Economic Inquiry, 59(3), 944-960.

Huntington-Klein, N., et al. (2025). The sources of researcher variation in economics (No. w33729). National Bureau of Economic Research.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science, 23(5), 524-532.

Knorr Cetina, K. (1999). Epistemic cultures: How the sciences make knowledge. Harvard University Press.

Kuhn, T. (1961). The function of measurement in modern physical science. Isis, 52(2), 161-193.

Kuhn, T. (1962). The structure of scientific revolutions. University of Chicago Press.

Kuhn, T. (1970a). The logic of discovery or psychology of research? Criticism and the growth of knowledge. Cambridge University Press.

Kuhn T. (1970b). Reflections on my critics. In Criticism and the growth of knowledge. Criticism and the growth of knowledge. Cambridge University Press.

Lakatos, I. (1978). The methodology of scientific research programmes. Cambridge University Press.

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Meta-Research: Investigating disagreement in the scientific literature. Elife, 10, e72737.

Latour, B., Salk, J., & Woolgar, S. (1979). Laboratory life: The construction of scientific facts. Princeton University Press.

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

Menkveld, A. J., et al. (2024). Nonstandard errors. The Journal of Finance, 79(3), 2339-2390.

Necker, S. (2014). Scientific misbehavior in economics. Research Policy, 43(10), 1747-1759.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. Elife, 5, e21451.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. Nature, 613(7942), 138-144.

Peterson, D. A. (2025). The Unbuilt Bench: Experimental Psychology on the Verge of Science. Columbia University Press.

Popper, K. (1959). The logic of scientific discovery. Routledge, London.

Popper, K. (1963). Conjectures and refutations: The growth of scientific knowledge. Routledge, London.

Schweinsberg, M., et al. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organizational Behavior and Human Decision Processes, 165, 228-249.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359-1366.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4(11), 1208-1214.

Simonton, D. K. (2004). Psychology's status as a scientific discipline: Its empirical placement within an implicit hierarchy of the sciences. Review of General Psychology, 8(1), 59-67.

Silberzahn, R., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science, 1(3), 337-356.

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. Social studies of science, 30(1), 73-94.

Solow, R. M. (1982). Does economics make progress?. Bulletin of the American Academy of Arts and Sciences, 36(3), 13-31.

Spescha, A. (2021). False feedback in economics: the case for replication. Routledge.

Starns, J. et al. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. Advances in Methods and Practices in Psychological Science, 2(4), 335-349.

Quine, W. (1951). Two dogmas of empiricism. In Can Theories be Refuted? Essays on the Duhem-Quine Thesis (pp. 41-64). Dordrecht: Springer Netherlands.

Woodward, J. (1989). Data and phenomena. Synthese, 393-472.

Weber, M. (2004). Philosophy of experimental biology. Cambridge University Press.

Yarkoni, T. (2022). The generalizability crisis. Behavioral and Brain Sciences, 45, e1.